Revisiting and Re-computing the X-Score Scoring Function

Hilaire Mobele Mambo (MMBHIL001)
University of Cape Town (UCT)

Supervised by: Prof. Jonathan Blackburn
University of Cape Town, South Africa

17 November 2014

Submitted in fulfillment of a Masters Degree in Bioinformatics at UCT

Abstract

Scoring functions seek to compute in different ways protein-ligand binding energies by summing together the individual pairwise atomic interaction energies observed in crystal structures between the protein and the bound ligand. To date though, accurate prediction remains a big challenge since existing scoring functions fail to reproduce known binding energies with a sufficient degree of accuracy and robustness. To overcome this problem, we assign a discrete weighting to the individual atomic interaction to account for entropic desolvation factors on ligand binding. We thereafter re-compute the revised scoring function and test the output against multiple sets of data to examine the robustness of the heuristic weightings used.

Declaration

I, the undersigned, hereby declare that the work on which this thesis/dissertation is based in my original work (expect where acknowledgments indicate otherwise) and that neither the whoole work nor any part of it has been, is being or is to be submitted for another degree or any other university. I authorise the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

signature removed

Signature.....

Hilaire Mobele Mambo,

Date : 17 November 2014

Contents

1. Introduction

1.1 Background

At the dawn of more personalised medicine, there remains a pressing need to understand in greater detail the effect of polymorphic variation and/or mutations in drug target and drug metabolising enzymes on drug efficacy. This is particularly important in the anti-cancer field where it is well known that drug target enzymes such as the epidermal growth factor receptor (EGFR) typically accumulate mutations during disease progression and that these accumulated mutations can confer either drug resistance or, paradoxically, drug sensitivity[Waters & MacLeod, 2003]. Today, however, the molecular cause of these effects are still rarely understood in any detail, even after the event, and our ability to predict such effects *a priori* remains in its infancy [Evans & McLeod, 2003]. In the absence of reliable computational approaches to this problem, there is still a very real need for, and reliance on, experimentally-derived quantitative thermodynamic data.

Purely experimental approaches to generate the requisite thermodynamic data on a multitude of protein-drug interactions (including binding constants, inhibition constants, rate constants, etc.) are hampered by the fact that few groups worldwide in industry or in academia - can handle the throughput of quantitative assays required across potentially hundreds of clinically-relevant variants, even for just a single target protein and a handful of drug-like molecules. There is therefore tremendous scope for the development of novel computational approaches that can accurately predict protein-drug interaction energies.

By way of example, if the binding affinities of a range of small molecule inhibitors (e.g. existing drugs and new drug candidates) could be accurately predicted across a range of clinically-relevant variants of the intended drug target, this information could then underpin drug differentiation data that, together with patient genotype, could be used by clinicians to inform the choice of the best drug for the specific drug target variant present in the individual patient [Society, 2005] i.e. personalised medicine!

Computational approaches to this problem are usually described as "scoring functions", of which a number are described in the literature and are available either as open source or commercial software packages [e.g. X-Score[Wang *et al.*, 1998]]. Essentially, scoring functions aim to compute in different ways the sum of the pairwise atomic interaction energies between a protein and a bound ligand; it is clear though that the existing scoring functions are not yet able to predict binding energies with sufficient accuracy or robustness.

In particular, the performance of typical scoring functions is limited in part by the simplistic assumption that pairwise atomic interactions are independent and in part by their reliance on Newtonian mechanics. In reality, protein-ligand molecular interactions are typically co-operative in nature, occur at the atomic level, and also involve significant solvation effects so can only really be modeled efficiently using quantum mechanics (QM). However, the computational resources required for such QM calculations on the thermodynamics of protein-ligand interactions are machine intensive and, whilst QM-based scoring functions represent a promising avenue for future research, they are not considered further here.

Taking EGFR kinase - a validated drug target in Non Small Cell Lung Cancer (NSCLC) as a case in point, the experimentally-derived *in vitro* inhibition constants of a number of EGFR kinase variants have been shown elsewhere to correlate reasonably well with patient survival in the clinic. For example, the *in vitro* inhibition by Iressa of the $L858R$ variant, the $G719S$ variant and the wild-type (Wt) EGFR enzyme follows the trend $L858R > G719S > Wt$, whilst clinical survival of patients having these

variants follows the same trend [[Paez, 2004], [Yun *et al.*, 2007]]. However, despite the existence of this experimental data on drug binding constants, preliminary calculations have shown that existing generalist scoring functions, such as X-Score, fail to accurately compute the binding affinity of drugs such as Gefitinib (aka Iressa), Erlotinib (aka Tarceva), Lapatinib etc. in their *in vitro* inhibition of the $L858R$, $G719S$ or Wt variants of EGFR kinase and even struggle to get the rank order of ligand binding strength. By extrapolation, there seems to be little prospect that such existing generalist scoring functions could currently accurately compute and rank the binding affinities of even this limited set of existing drugs against the $>$100 known clinical variants of EGFR for which experimental data is not yet known.

The present work therefore seeks to develop a novel heuristic computational approach to solve this problem, with the aim of creating an accurately predictive computational model of interaction affinities between protein kinase variants and drug-like molecules.

In particular, the hypothesis explored in this Thesis is that, by assigning heuristic weightings to individual pairwise atomic interactions, it may be possible to effectively allow for differential desolvation factors that affect protein-ligand interactions and which are currently not accounted for in scoring functions such as X-Score, thereby obtaining a better and more robust correlation between computed and experimental binding energies. The case study used to explore this hypothesis here is based on the performance of the X-Score scoring function on protein kinases; the results and discussion will focus thereafter on the sensitivity and re-parameterisation of X-Score.

2. Overview of Scoring functions

2.1 Introduction

In Medicinal Chemistry, the likely potency of a drug candidate is often assessed initially by its association constant with the protein receptor, which is in turn related to the free energy of binding. There are currently a number of ways to estimate this value computationally of which the most physically realistic are techniques involving methods such as molecular dynamics (MD) or Monte Carlo (MC) simulation, and statistical mechanics tools such as free energy perturbation (FEP) [Kollman, 1993] that extracts the relevant information from the simulation trajectories. However, these methods are very complex and computationally expensive and, even with the modern availability to cheap super computing power, require running times in the order of days per ligand.

Docking programs use scoring functions to return the predicted free energy of each identified three dimensional structure by estimating the interaction energy between ligand and receptor. A docking program uses scoring functions in two phases. The first phase tries to identify the true mode of binding for the ligand; To achieve this each molecule is associated with an ensemble of conformations and these are then posed inside the receptor in different orientations. The energy returned by scoring function is kept if the score is good and dropped if it is not. In the second phase, poses are now ranked according to their computed scores. The scoring functions used to find the various orientation and to rank the poses are not necessarily the same. To obtain a correct ranking that correlates with experimental free energies of binding, a more elaborate scoring functions is required to identify the true binding mode [Jhoti & Leach, 2007].

2.2 Parameters in scoring functions

The components required for scoring functions can be developed by considering the fundamental physics of intermolecular interactions. The accurate evaluation of electrostatic interactions is one of the hurdles in binding free energy and is modeled using the Coulomb's law. Hydrogen bonds are still assessed by electrostatic interactions between an electronegative atom (acceptor) and a hydrogen atom covalently bound to another electronegative atom (donor). Hydrogen bonding gives some specificity to the binding process because all hydrogen bonding sites in a protein-ligand complex should be satisfied for optimum binding to be observed[Bohm & Klebe, 1996]. Van der Waals interactions are also another major contributing factor to interaction energies and parameterise the interaction between nonpolar molecules, giving balance between attractive and dispersion forces [[Atkins, 1998], [Ajay & Murcko, 1995]]. This hydrophobic effect explains the preferential association between non-polar molecules, or an area of molecules, to minimize water contact. The above list of parameters in scoring functions is far from exhaustive since the calculation of free energy of binding is not fully understood yet. Moreover, the above parameters only attempt to describe the enthalpic component of a ligand binding event, not the enthropic component which is itself to be at least as important (according to the second law of thermodynamics). The enthropic component is generally broken down into :

1. translational entropy,
2. rotational entropy,

3. vibrational entropy.

Of these, translational and rotational entropy are the biggest factors in ligand binding but are difficult to accurately compute for large molecules.

2.2.1 Binding Affinity

The interaction between the Protein - Ligand can be mathematically formalize by the equation below:

$$RL \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} R + L \tag{2.1}$$

where R is the receptor, L the ligand, RL product or the complex, k_{-1} is the association rate constant for the reaction going from right to left, and k_{+1} is the dissociation rate constant for the reaction going from left to right [Ajay & Murcko, 1995].

So the equilibrium constant (the dissociation equilibrium constant) is defined as

$$K_d = \frac{[R][L]}{[RL]} \tag{2.2}$$

where K_d is the equilibrium constant used for expressing binding affinity and is usually referred to as the dissociation constant. When the value of K_d is small, the tendency of RL to dissociate is small, i.e., RL tends to remain as a complex.

The thermodynamic equation that relates the free energy change to change in enthalpy and entropy is given by the equation,

$$\Delta G = \Delta H - T\Delta S \tag{2.3}$$

where ΔG denotes the change in free energy of reaction, ΔH and ΔS are the corresponding changes in enthalpy and entropy, and T is the temperature of the system. The equilibrium constant is related to the free energy change [Ajay & Murcko, 1995] of the dissociation of RL as,

$$\Delta G = \Delta G^{\circ} + RT \ln K_d \tag{2.4}$$

Where, R is the gas constant, and T is the absolute temperature. ΔG° is free energy change associated with the reaction under standard conditions. At equilibrium $\Delta G = 0$, so that we have,

$$\Delta G^{\circ} = -RT \ln K_d \tag{2.5}$$

The K_d for model system is only a function of temperature, this is why it is therefore called the equilibrium constant.

2.2.2 Binding constant, dissociation constant and $IC50$

The inhibition constant K_i is measured in kinetic experiments where both the substrate and the inhibitor are present, but not necessarily at equilibrium unless steady state kinetics is assumed. More direct K_i measurements (usually called a K_d measurement) that do not involve competition with substrate are equilibrium measurements. All comparison with experimental data should therefore take place when the experiments are performed under equilibrium conditions [Ajay & Murcko, 1995].

The interaction of ligand with a protein may also be measured in terms of the $IC50$ value. The $IC50$ represents the concentration of a drug or inhibitor that is required to reduce the binding of a ligand (or rate of reaction) by half. In theory K_i or K_d is preferred to $IC50$ because $IC50$ depends on the amount of ligand available to the receptor and therefore the comparison of data obtained under different conditions is impossible. The binding constants K_i or K_d can be compared more easily. In principle, for any inhibitor $K_i = K_d$. Determination of binding constants require more data than is needed for $IC50$. The $IC50$ is then defined as the concentration of ligand at which 50% of the receptor sites are occupied in one to one complex. As a result, $IC50$ values are not true inhibition constants. Another measure associated with drug potency is $EC50$ or half maximum effective. It is the concentration of a drug, antibody or toxicant required to obtaining 50% of the maximum effect after some specified exposure time. Like $IC50$values, $EC50$ values are therefore not true inhibition constants [Ajay & Murcko, 1995].

2.2.3 Thermodynamics of Binding

The difference in free energy between the bound and the free states is the quantity of interest for determining the binding constants, given by

$$\Delta G = G^c - G^u \tag{2.6}$$

The superscript "c" and "u" are for complexed and uncomplexed, respectively. So, G^c and G^u means the total free energy of all interactions in bound (i.e complex) and unbound or free states respectively.

As stated previously, the calculation of free energy of binding is not an exact science yet and different approaches to scoring have led to different kinds of scoring functions. They are usually classified in a tree-like structure, with three principal branches, plus one "hybrid" type. The three principle types force-field, empirical, and knowledge-based scoring functions as well as the "hybride type", consensus scoring will be discussed below [Halperin *et al.*, 2002].

2.3 Force Field scoring functions

This class of scoring function uses the molecular force fields where the vibrational potential energy of a diatomic molecule can be approximated using Hook's law for a spring. The Hook's equation is given by :

$$U(r) = \frac{1}{2} k \Delta r \tag{2.7}$$

Where k is similar for similar bonds (C - C, C - N, C - S, etc.). Force fields are used to provide this information.

Among the parameters used to develop force fields are bond stretching, valence angle bending, torsional angles, Van der Waals interactions and electrostatic interactions. The force fields also take into account the desolvation effect and the action of water as a solvant [Shoichet *et al.*, 1999] using either a distance dependent dielectric constant [Vieth *et al.*, 1998] or a complete Poisson-Boltzmann treatment. The computation of Poisson-Boltzmann term is very expensive so it is typically substituted by other faster approaches like the generalized-Born model(GB/SA). Usually the algorithm used takes approximately 10 seconds per orientation on a silicon graphic R10000 computer [Zou *et al.*, 1999].

The force field scoring functions also take into account the internal conformational energy of the ligand. It is generally accepted that when a ligand binds inside a binding site, it will do so in a low energy conformation [Bostrom *et al.*, 1998]. The components of the Amber [Cornell *et al.*, 1995], Charm[Brooks *et al.*, 1983] and MMFF(Merck Molecular FF) [Halgren, 1996] are used as scoring functions in many docking programs.

In recent year Pearlman and Clarifson introduced a new approach to force field scoring [Pearlman, 1999] [Pearlman & Charifson, 2001] in which they developed and tested OWFEG (**O**ne-**W**indow **F**ree **E**nergy **G**rid), which is a grid approximation of the free-energy perturbation method. In **OWFEG**, a short Molecular Dynamics (MD) simulation is used to build a grid of free energy values relatives to some probes (neutral, charged, methyl) around the active site. The score is obtained by linear interpolation of these values for a given docked pose and K_i value can be predicted [Pearlman, 1999] using this method.

2.3.1 Problems and solutions

The main drawback with force-field scoring functions is that their computation is quite time consuming. Furthermore, parameters that are initially derived for other simulation methods, such as MD or Monte Carlo (MC) calculations, may not be suited for single-point energy estimates; correlations with experimental data may be poor, especially when the ranked compound does not belong to the same class. Poor charge modeling also seems to have a detrimental effect. The obvious solution to these problems seems to be the development of potentials and charge models for specific use in scoring functions. Another remedy could be the introduction of new non-bonded terms to estimate interactions that until now have been neglected or merged with others [Fenu *et al.*, 2007]. As an example[[Raha, 2005] [Raha & Merz, 2005] [Raha & Merz, 2004a] [Raha & Merz, 2004b]] we can cite a quantum-mechanically-based semi-empirical Hamiltonian approach to scoring functions used to estimate the metal-ligand interaction contribution to free energy of binding. Although currently too slow for virtual screening, there is clear opportunity for such methods to be more widely used if they provide information that simpler scoring functions cannot.

2.4 Empirical Scoring functions

Another class of scoring functions is the empirical scoring functions. The idea behind these scoring functions is that the free energy of binding is split into series of terms, each of which has a specific and intuitive explanation. It turns out however that the assumption of decomposing the free energy of binding into different components is not accurate because the free energy of binding is a function of state whereas the energy components into which it may be divided are not [Mark & Gusteren, 1994]. Empirical Scoring functions do lead however to an interesting approximation that proves to be useful to some extent. The computational approach to calculation of these terms is typically relatively simple and many terms can therefore be used and implemented in different ways. Such empirical scoring

methods are used for example in structure-based virtual screening (SVS). The first attempt at empirical scoring involved the building of a knowledge base from the analysis of group-to-group interaction in 200 complexes with known experimental binding affinity [Andrews *et al.*, 1984]. These methods do not use the three-dimensional structure of the complex, as they simply count the number of interactions that were presumed to exist in the complexes used as reference.

Later, atom-to-atom approaches were introduced, with the knowledge of the three-dimensional structures of the training set leading to a greater level of detail in the analysis of interactions between receptor and ligand [Bohm & Klebe, 1996]. The mathematical expression of the free energy of binding is given as a sum of interaction terms, each multiplied by its own weight or coefficient. The formula is written as :

$$\Delta G = \Sigma_i f_i \Delta G_i \tag{2.8}$$

Where, f_i are the coefficients, and ΔG_i are the free energies associated with each term, each term reflecting an interaction involved in the binding process. Other non-linear functional forms have been investigated [Giordanetto *et al.*, 2004], but results do not seem to be very conclusive as non-linear terms do not sufficiently improve predictive power to justify the computational overhead and loss of physical meaning on moving away from familiar terms. Coefficients are generally obtained through optimisation/regression techniques in which the scores given by an equation are fitted to known experimental binding affinities. The work by Smith et al. [Smith *et al.*, 2003] provides a different approach where the scoring function is instead parameterised using enrichment of true binders interspersed among a number of decoy molecules.

2.4.1 Parameters in empirical scoring functions

Common terms in empirical scoring functions include those for hydrogen bonds, ionic interactions, hydrophobic interactions, and the internal energy of the ligand. The parameters in empirical scoring functions are similar to non-bonded force field potentials where parameters are provided by fitting score to known experimental binding affinities. Because of the additive feature of empirical scoring functions, larger ligands tend to score better than small ones. However, larger ligands suffer from the fact that on binding, more degrees of freedom are frozen by confinement in the receptor pocket and the consequential the entropy change usually disfavours binding. Terms that try to estimate this entropy change, predominantly through rotatable bond counts or similar quick methods, are therefore typically added. Another approach to account for ligand size is to scale the scores obtained, for example Pan *et al.* suggested multiplying the score by the square root of the number of heavy atoms, or the ligand's molecular weight [Pan *et al.*, 2003], but this seems a rather crude and overly simplistic approach and is not widely used.

2.4.2 Problems and Solutions

Despite their success in modeling some complex interactions with relatively simple equations, empirical scoring functions have their limitations. A major drawback of any regression-based scoring function is the dependence on the size, composition and generality of the training set used to derive the weights [Fenu *et al.*, 2007]. To counter this, efforts are being made to create common training and validation sets, containing diverse proteins and ligands. Another common approach is to build scoring

functions 'tailored' for a certain class of protein or ligand [Bohm, 1994]. A further problem with empirical scoring functions is that they, like force field scoring functions, often lack terms to account for some interactions. For example, metal-ion interactions are often neglected or approximated by simple coulombic terms, without any reference to charge transfer and coordination. Unfortunately, gathering sufficient information to parameterise such terms can be difficult, so their contribution to the binding is often merged into other terms which can create obvious problems.. For example, if metal ion interactions are neglected and we do not have proteins that use metal to bind the ligand in our training set, then the scoring function will obviously fail when applied to such a protein. Alternatively, if we have a few cases in our training set, but no term to account for metal binding, then regression analysis may try to compensate by increasing the factor for the coulombic interaction and even in the case of metal-proteins, this function would not behave correctly, as a modest over estimation would not be sufficient to address the missing metal ion term. The answer, in this case, is to develop and add new terms and once again to select accurately the training set to reflect the application field. Another problem inherent in the empirical scoring function approach is that the training sets used to drive such functions usually do not contain "negative" data (i.e unfavourable inter- or intra- molecular interactions) since these are very rarely observed in X-ray structures. Finally, it is of the utmost importance to remember that these models are just that, and as such will always be limited in scope and applicability [Fenu *et al.*, 2007].

2.4.3 Some Advances

A number of novel empirical scoring functions have appeared recently some of which are simple re-editions of earlier functions. For example, Verdonk *et al.* [Verdonk *et al.*, 2003] re-implemented ChemScore [Eldridge *et al.*, 1997] within GOLD [Jones *et al.*, 1997] and conducted studies on the influence of using different scoring functions for the sampling and ranking parts of the docking process using the CCDC/Astex docking test set divided into fragment-like and drug-like ligands. Overall, this study confirmed that using GoldScore for both sampling and ranking yields more accurate results; this is in contrast with previous assumptions that using two different scoring functions for the two docking phases would improve the quality of the poses (for example, using ChemScore for sampling and the outcomes (results) being re-ranked with Goldscore). Verdonk *et al* [Verdonk *et al.*, 2003], allowed for water molecules to be switched on and off at their experimentally determined position during docking; GOLD's GoldScore and ChemScore functions were then modified to consider the loss of water's rigid body entropy upon binding, with a constant penalty term σ^p, obtained from a training set of 58 complexes, and tested against 225 complexes. The method managed to correctly predict water displacement/mediation for over 90% of the complexes in both training and test sets.

Mancera *et al.* [Mancera *et al.*, 2004], in their program EasyDock [Todorov *et al.*, 2002], used PLP [Gehlhaar *et al.*, 1995] to carry out sampling, and ScreenScore, which put together the functionality of PLP and FlexX's own score, to re-rank the results [Stahl & Rarey, 2001]. The program GEMDOCK, [Yang & Chen, 2001] uses a generic evolutionary method for docking and a novel empirical scoring function built with three terms: coulombic interactions with simplified charges; PLP-like hydrogen bond terms; and internal energy [Fenu *et al.*, 2007]. This final term is in fact a constant penalty to maintain poses to a pre-defined docking box. The program has been tested for binding-mode recognition against 100 protein/ligand complexes selected from the Protein Data Bank (PDB), the results of which show that in 79% of these complexes, the docked lowest energy ligand structures had root-mean-square deviations (RMSDs) less than $2.0\mathring{A}$ with respect to the corresponding crystal structures. The success rate increased to 85% if the structure water molecules were retained. GEMDOCK was evaluated on two cross-docking experiment in which each ligand of the set of protein was docked into each protein

of the group. 76% of the docked structures had RMSDs below $2.0\mathring{A}$ when the ligands were docked into foreign structures. The analysis and validation from GEMDOCK performance with respect to various search spaces and scoring functions, shows that the outcome is function of the quality of scoring function, if the degree of accuracy of the scoring function used is high , then the predicted accuracy from GEMDOCK will also be high. The authors then conclude from their finding that GEMDOCK is a useful tool for molecular recognition and may be used to systematically evaluate and improve scoring functions [Stahl & Rarey, 2001].

A new version of GLIDE has also been described [[Friesner *et al.*, 2004] [Halgren *et al.*, 2004] which uses a series of filters of increasing complexity to select good binding modes and includes a scoring function derived from ChemScore [Eldridge *et al.*, 1997], extended with a few more terms. According to the authors the overall performance of GLIDE is equal or superior to those of GOLD and FlexX.

An opportunity to improve the performance of scoring functions lies in the method by which a function is parametrised. Smith *et al.* [Smith *et al.*, 2003] have built a novel scoring function by selecting terms from Gschwend's TEC toolkit containing a variety of 2 and 3-D descriptors for protein-ligands interactions. The novelty of the method does not lie in the terms selected like hydrogen bonds, coulombic interaction, a contact/clash count, and buried hydrophobic surfaces term; but instead the parametrisation method attempts to mimic a Structure-based Virtual Screening (SVS) experiment, with a training set of twenty proteins and 1000 ligands selected from the World Drug Index (WDI). The ligands were used as decoy to hide the known ligand, while a genetic algorithm adjusted the parameters to optimise the enrichment factors. The genetic algorithm (GA) is an optimization and search technique based on the principles of genetics and natural selection. A GA allows a population composed of many individuals to evolve under specified selection rule to a state that maximizes the fitness [Haupt & Haupt, 2004].

2.5 Knowledge-based scoring functions

This class of scoring functions has been developed, thanks to the huge increase in crystal structure with high resolution obtained by X-ray diffraction, solid state NMR and a number of other related methods. The fundamental idea of a knowledge-based scoring function is"to extract statistical information about the ligand/protein binding modes and to correlate these to free energy of binding using statistical mechanics. Knowledge based pseudo-potentials are usually derived by accumulating radial distribution functions for selected pairs of protein-ligand atoms. The inverse formulation of the Boltzmann equation, or a derivative thereof, is then used to extract the corresponding potentials of mean force from these distributions. To apply these potentials to scoring, the distances between the protein and ligand atoms in each pose is calculated and the total score is obtained by summing each of the energies derived from the potential curves using these distances"[Fenu *et al.*, 2007]. The Boltzmann equation can be written as:

$$E_{ij} = -kT \ln \rho_{ij}(r) - kT \ln Z \tag{2.9}$$

Where, E_{ij} is the interactions energy involving atom i on the protein and atom j on the ligand; ρ_{ij} is the radial distribution function between the two atom-types at distance r and Z refers to partition function. The fact that no particular function has been imposed to form the pseudo-potential curves, it is implicitly accepted that any possible kind of interaction could occur, irrespective of whether it is enthalpic or entropic [Fenu *et al.*, 2007].

It must also be noted that besides the functional form used to derive the potentials from the populations, the differences between various incarnations of knowledge based functions lie in the amount of data used for parametrisation and in the number of atom types considered. It has been suggested that the more protein-ligand complexes used in parameterisation, the better as higher quality statistical information can be extracted from a larger knowledge-base. This consideration is also relevant to the the diversity of the chosen parameterisation set, which will produce a much more robust scoring function. On the other hand, increasing the number of atom types, does not necessarily build a better potential: Add too many and the data becomes fragmented because there may not be sufficient information to build pseudo-potential curves between certain atom pairs, with the data effectively being wasted plus there being a risk of over-fitting the function. Conversely, if there are insufficient atom-types, different interactions will be incorrectly grouped, giving potential curves that, although smoother, may have unphysical features: For example, a single pseudo potential curve with two minima could arise from superposition of two single-minima curves. Alternatively, it may signify the presence of some angular-dependent phenomena, which radially-symmetric potentials cannot incorporate effectively [Fenu *et al.*, 2007]. Good practice is arguably to choose well-known and well-defined atom-types, as Gohlke and coworkers did with DrugScore [Gohlke *et al.*, 2000], which uses 17 atom-types taken from the Sybyl mol2 format. This appear to be a good choice as it reduces the difficulties of reliable ligand atom-typing. Others such as Muegge and Martin [Muegge & Martin, 1999] extract a potential with more than 17 atom types used to represent the diversity of pairwise interaction in their parametrisation set, or adopted custom-defined atom. In the improved Potential Mean Force (PMF) scoring function [Muegge & Martin, 1999], they choose atom-types based on chemical features and not the atom's hybridisation state.

The first applications of knowledge-based scoring functions to drug design were strictly focused on HIV protease [[Mizutani *et al.*, 1994], [Verkhivker *et al.*, 1995]] the only proteins at the time for which there were sufficient structural data to develop a scoring function. The outcomes was promising, the early attempts did not return a generally applicable function. Among other programs utilising a knowledge-based pseudo-potential, we can cite are SMoG [[DeWitte & Shakhnovich, 1996], [DeWitte *et al.*, 1997] and BLEEP [Mitchell *et al.*, 1999] . In the latter, functional form for the shell density is different, and the knowledge-based function is complemented by a Van der Waals term to compensate for the low occurrence of short range interactions in crystals. The PMF scoring function [Muegge & Martin, 1999] also contains a volume correction factor, to account for the volume around each ligand atom occupied by other ligand atoms. This function was tested against the FKBP protein [Muegge & Martin, 1999] and it was modified to improve virtual screening performance by substituting the knowledge-based curves with Van der Waals terms to account for short-range steric effects otherwise neglected. In DrugScore [Gohlke *et al.*, 2000], a knowledge-based one-body potential, scaled to the size of the solvent-accessible surface (SAS) of the protein and the ligand that becomes buried upon complex formation, was added to PMF-like pairwise pseudo-potentials. This results in DrugScore effectively being a mixed knowledge/empirical scoring function. Ishchenko *et al.* released a new version [Ishchenko & Shakhnovich, 2001] of SMoG (SMoG2001), parametrised on 725 structures, and tested against 119, reporting performance similar to DrugScore and better than PMF and SCORE1 (LUDI [Bohm, 1994], the improvement being mainly attributed to a better description of the reference state used to normalise the radial distribution function.

The biggest drawback of knowledge-based scoring functions, apart from doubts about their statistical-mechanical meaningfulness, is that not all interactions can be approximated efficiently by pairwise terms. Moreover, many interactions are highly directional, whereas radial distribution function-derived pseudo-potentials have spherical symmetry. Although a certain degree of directionality comes from the interplay of different pairwise interactions, the issue has not yet been fully addressed. Despite this limitation, these functions are cheap to evaluate and work reasonably well within the lead-identification

phase of drug-design. With the constant increase in the number of resolved protein structures, the reliability and general applicability of these functions is likely to improve, and the possibility (re-)opens for functions to be optimised on certain protein-ligand classes, or even for different alternative non-pairwise representations of the intermolecular interactions [Fenu *et al.*, 2007].

2.6 Consensus scoring

Since none of the scoring function classes presented above is clearly superior to others, a more pragmatic approach to obtain good results is to combine their separate judgments. This approach, called "consensus scoring", was pioneered by Charifson *et al* [Charifson *et al.*, 1999]; two or more scoring functions are applied to the same set of poses and only structures that perform well in more than a pre-defined share of the said scoring functions are retained. Consensus scoring, analyzed in detail by Wang and Wang [Wang & Wang, 2001.], is effective in reducing the number of false positives, which may be able to trick one function, but not all, into believing they are actives. For this reason, performance is particularly good when functions included in the consensus are different, in the sense that they describe different aspects of binding. For example, a scoring function oriented towards the careful detection of hydrogen bonds will complement a more hydrophobic function, whereas the union with a more similar function may not be so productive. An example of consensus scoring function is ScreenScore [Stahl & Rarey, 2001], built from the consensus of PLP and FlexX. In many cases however, the description of the protein-ligand interactions is still deficient, so there remains room for new terms to be added.

2.6.1 Recent Advances

The consensus scoring field has been very active recently, with a number of interesting papers exploring new solutions and extending and/or confirming other researchers' ideas and suggestions. Verdonk *et al* [Verdonk *et al.*, 2003] tested a number of different consensus combinations and confirmed Wang's insight that rank-by-number outperforms rank-by-rank and rank-by-vote strategies in consensus scoring. Wang [Wan, n.d.] and colleagues again, using three scoring functions built upon their previous work (SCORE [Wang *et al.*, 1998]), creating a new consensus scoring method (X-SCORE) which was calibrated to reproduce the binding affinities of 200 complexes, and tested against thirty more.

Guo et al. [Guo *et al.*, 2004] built a consensus scoring function within Sybyl consisting of ChemScore, G-Score, F-Score, PMF-score and DrugScore. These five were combined using multi-linear regression. The training set here consisted of 53 inhibitors of *Torpedo Californica* AChE from PDB entries with known affinity re-docked into a human AChE and 16 compounds were used as a test set.

In the same fashion, Jacobsson [Jacobsson *et al.*, 2003] applied a number of different training methods to 'multidimensional' consensus scoring with partial least squares, discriminant analysis, Bayesian classification, and rule-based methods being applied to improve discrimination between active and inactive compounds.

Klon *et al.* [Klon *et al.*, 2004] used a naive Bayesian machine-learning algorithm to improve enrichment in high-throughput docking of databases by selecting the important features from the top ranked structures with extended connectivity fingerprints and correlating these to their high-scoring compounds. This method works by re-ranking the docking outcome and retrieving compounds similar to those at/or near the top of the list from the remainder.

2.7 Comparisons between scoring functions

There is a growing interest in comparing different scoring function head to head to identify common deficiencies and improve scoring methods. There are a number of recent papers that tackle this issue from different view points. Wang *et al.* [Wang *et al.*, 1998] tested 11 freely and commercially available scoring functions by reproducing the affinity of 100 protein-ligand complexes. Later, the same authors extended the test, by using 14 different scoring functions and binding affinities of 800 complexes from the PDBBind [Wang *et al.*, 2004] database. Ferrara *et al.* [Ferara *et al.*, 2004] assessed 9 scoring functions for their ability to recognize the correct ligand orientation in 189 complexes from LigandPDB [Roche *et al.*, 2001]. Krovat *et al.* [Krovat & Langer, 2004] applied LigandFit to the test case of renin : 10 inhibitors were mixed with 990 drug-like compounds, docked and then analysed with seven scoring functions: LigScore1, LigScore2, PLP1, PLP2, JAIN, PMF, LUDI. The consensus combination PLP1 + PLP2 + PMF performed the best, presenting all known inhibitors within the top 8%.

In another insightful study Kontoyianni *et al.* investigated the sensitivity of five docking programs (FlexX, GOLD, DOCK, LigandFit, GLIDE) to the nature of the active site. 69 receptors (belonging to 14 families) were used as targets. Their findings show that GOLD, followed by Glide, perform well in predicting accurate poses. In the same fashion, Perola *et al.* [Perola *et al.*, 2004] compares three docking programs GLIDE, GOLD and ICM, to identify correct binding mode using a mixed test set of Vertex in house complexes and complexes from PDB data bank. The objective of the study was to assess how the nature of active site is influenced and what is the effect of energy minimization on scoring. Glide was able to correctly identified the crystallographic pose within $2.0\mathring{A}$ in 61% of the cases, against 48% for GOLD and 45% for ICM. In most of the case the performance of the Glide appears to be consistent with respect to diversity of binding sites and ligand flexibility, while the performance of ICM and GOLD is likely dependent on the binding site and it is significantly poorer when binding is done mainly by hydrophobic interactions. The findings also show that energy minimization and re-ranking of the top N poses can be an effective means to overcome some of the limitations of a given docking function. Another test compares these three programs on HIV-1 protease, p38 MAP Kinase and IMPDH, which reveals that the performance of GLIDE is better than the other two programs on these target. DOCK, FlexX, GLIDE, GOLD, Slide, Surflex, and QXP have been tested by Kellenberger *et al.* [Kellenberge *et al.*, 2004] in a 100 X-ray small-ligand structure reproduction and simulated SVS experiment. As result of the test GOLD, GLIDE, and Surflex did well at both tasks, but the performance of the docking programs was subject to criticism by the authors on the basis of their active features, and the common failures that were identified. In Vigers *et al.* [Vigers & Rizzi, 2004], once again, these studies show that FlexX and Gold are able to correctly select the right ligand orientation, but fail to consistently rank a series of ligands. A statistical multiples active correction (MASC) was presented which reduces for ligands that are found to score well not only with the target of interest, but also with another selected set of seven to nine diverse protein structures. This correction was shown to improve ranking, but unfortunately it is specific to the docking and scoring algorithm used, and so must be re-generated for any new combination [Fenu *et al.*, 2007].

2.8 New Test and Training Sets

Many scoring functions have been calibrated with different data sets making the comparison between functions difficult. As a result of this, the notion of having a common training set for all functions has been suggested by researchers in the field. It is generally recommended that the training sets should

include many diverse proteins so it is representative of the 'protein space'. The ligands to which the protein are complexed should also be diverse and arguably should all be drug-like [Lipinski *et al.*, 2001]. Numerous attempts have been made to provide standard training and test sets, the most comprehensive contribution to date coming from PDBbind by Wang *et al.* [Wang *et al.*, 2004]. The new version of PDBbind, released in 2012, provides binding affinity data for a total of 9308 biomolecular complexes in the PDB, including protein-ligand (7121), nucleic acid-ligand (79), protein-nucleic acid (511), and protein-protein complexes (1597). The Mother of all Data Base (MOAD) or Binding MOAD, whose goal is to provide the largest collection of well resolved protein crystal structures (2.5 Å or better) with experimentally determined binding data extracted from literature, currently contains 18,764 Protein-Ligand structures; 6,311 structures with Binding Data; 9,048 different ligands [Hu *et al.*, 2005]. The BindingDB [Liu *et al.*, n.d.] is another source of measured binding affinities containing 910,836 binding data for 6,263 protein targets and 378,980 small molecules.

In this effort to standardize the data sets there is still challenge of reliable experimental data among which we can list :

1. Different databases report different values for$K_i/K_d/K_n/IC50$ for the same complexes,
2. This reflect the fact that different labs use different experimental methods and different models to generate and interpret binding data,
3. All of these of course impact strongly on the performance of any empirical scoring function, as well as on the assessment of relative performance of all scoring function.

2.9 Summary

In this Chapter, a broad review of the scoring functions in the context of docking small molecules to protein targets has been presented. The basics of scoring functions have been presented as well as the recent advances in each three classes of scoring functions. We also discussed consensus scoring, comparisons between scoring functions and finally the collection of structural data with the specific purpose of scoring function testing. Our next Chapter focuses on X-Score, one of the best performing empirical scoring functions integrated into the molecular design tool kit for drug discovery and docking.

2.10 Problem statement

Many workers in the field of Medicinal Chemistry have appreciated the limited ability of existing scoring functions to distinguish correctly between protein-ligand binding strengths and a number such groups are tackling this problem through the use of more sophisticated free energy calculations.

However, in our opinion the weakness of existing scoring functions is only in part on their reliance on Newtonian mechanics to calculate the pairwise atomic interaction energies; instead, we believe that their major weakness lies in their failure to allow for the entropic factors arising from desolvation effects and protein conformational changes on ligand binding. Free energy calculations, which try to address these issues are however very complex and computationally expensive. We therefore seek to develop a simpler method that addresses these limitations.

In other words, we believe that the underlying cause for the failure of existing generalist scoring functions to recapitulate experimentally-determined protein-ligand binding trends is because they fundamentally assume that the enthalpic component of the protein-ligand interaction in the ligand-bound state dominates the binding affinity and because they also assume that the amino acid residues of the target proteins make equal contributions to drug-target binding strength, both of which are limiting assumptions.

We therefore propose to take a different, heuristic approach to the problem, based not on revamping the pairwise atomic interaction energy calculations themselves, but by assigning weightings to the individual pairwise atomic interactions. Thereafter, we re-compute the revised scoring function and test the output against real experimental data to examine the robustness of the weightings used.

2.11 Aim of thesis

The aim of this thesis were :

1. To test the sensitivity of the X-Score scoring function with the kinase/protein family.
2. To evaluate the performance of the differential weighting on hydrogen-bonding in interactions in X-Score against experimental data sets in order to assess the robustness of a heuristic approach to desolvation problem.

3. X-Score

3.1 Introduction

X-Score is a general empirical scoring function for estimating the binding affinity of a protein-ligand complex. It has major potential for application in structure-based drug design studies. There are three individual empirical scoring functions implemented in X-Score, which are named as **HPScore**, **HMScore** and **HSScore**, respectively. They can be conceptually summarized [Wang *et al.*, 1998] as:

$$HPScore = C_{0,1} + C_{VDW,1} \times (VDW) + C_{HB,1} \times (H-Bond) + C_{HP} \times (HP) + C_{R,1} \times (Rotor) \tag{3.1}$$

$$HMScore = C_{0,2} + C_{VDW,2} \times (VDW) + C_{HB,2} \times (H-Bond) + C_{HM} \times (HM) + C_{R,2} \times (Rotor) \tag{3.2}$$

$$HSScore = C_{0,3} + C_{VDW,3} \times (VDW) + C_{HB,3} \times (H-Bond) + C_{HS} \times (HS) + C_{R,3} \times (Rotor) \tag{3.3}$$

Here VDW, H-Bond, HP, HS account for Van der Waals contacts, Hydrogen Bonding, Hydrophobic Pair, Hydrophobic Match, Hydrophobic Surface respectively upon the binding process. The constants $C_{0,i}$, $C_{VDW,i}$, $C_{HB,i}$, $C_{R,i}$ with i = 1, 2, 3 and C_{HP}, C_{HM} and C_{HM} are coefficients of different terms that were obtained from the three linear equation above through regressional analusis on a training set composed of 200 protein-ligand complexes. The goal for developing X-Score was [Wang *et al.*, 1998]:

1. Have a fast, accurate and robust scoring function for structure based drug design;
2. Provide a practical tool to interpret the interaction between a ligand and its target protein.

The underlying idea in the X-Score method is that binding affinity is decomposed into the contribution of individual atoms. Each atom in a ligand has what is called an atomic score, indicating its role in the binding process. This major innovation of atomic binding allows those who work in drug discovery, for example, to inspect and optimize the lead compound structure in a more rational way [Wang *et al.*, 1998]. Below, we describe the method or algorithm used to compute each term in the function.

3.2 Methods and Algorithm

3.2.1 Training set

X-Score was calibrated with a training set comprised of 170 protein-ligand complexes. All the complexes came from the PDB (Protein Data Bank)archive and the data included more than 17 different proteins types, with end structure having a resolution better than 3.2 Å (Ångtröm), making X-Score a generalist scoring function. All the experimental binding was taken from literature and expressed as the negative logarithm of dissociation equilibrium constants, pK_d [Wang *et al.*, 1998].

The authors of X-Score used the SYBYL software to prepare the receptor and ligand for analysis. To do so, the first extracted the ligand from the complex structure and assigned proper atom type and bond type to the ligand in a new mol2 file. The protein was written to a separate file in PDB format; any molecules such as water molecule, metal ions, and other co-factor were treated as part of the protein. Protons were also added to the protein and ligand [Wang *et al.*, 1998].

3.2.2 X-Score Scoring Function Algorithm

The empirical scoring functions are built on the principle that the free energy change in the protein-ligand binding process can be dissected into basic components. The function equation in X-Score takes the following form :

$$\triangle G = \triangle G_{vdw} + \triangle G_{H-bond} + \triangle G_{deformation} + \triangle G_{hydrophobic} + \triangle G_0 \tag{3.4}$$

Where $\triangle G_{vdw}$ denote the contribution of Van der Waals contacts between the protein and its ligand; $\triangle G_{H-bond}$ represents the hydrogen bonding between the ligand and the protein; $\triangle G_{deformation}$ accounts for the deformation effect; $\triangle G_{hydrophobic}$ specifies the hydrophobic effect; $\triangle G_0$ is the regression constant that may contain the translation and rotational entropy loss that occurs on bonding [Wang *et al.*, 1998].

Each term in the function is computed as follows :

Van der Waals (VDW) interaction. The Van der Waals interactions are balance between dispersion forces and short range repulsion and they play a fundamental role in binding processes, but there is still disagreement on the best method to calculate them. X-score in its original approaches computed VDW by simply pairwise counting the VDW bumps between protein and the ligand which simply means that by summing out Van der Waals radii of the two interacting atoms (i.e atom from protein and Ligand).

In later version of X-Score uses Lennard Jones potential to reflect the balance between the short-range repulsion and the long-range attractive dispersion force [Wang & Wang, 2001.]. Although there exist many version of the Lennard Jones potential X-Score uses the so-called 8-4 version which can been mathematically written as :

$$\begin{aligned} VDW &= \sum_{i}^{ligand} \sum_{i}^{protein} VDW_{ij} \\ &= \sum_{i}^{ligand} \sum_{j}^{protein} \left[\left(\frac{d_{ij,0}}{d_{ij}} \right)^8 - 2 \times \left(\frac{d_{ij,0}}{d_{ij}} \right)^4 \right] \end{aligned} \tag{3.5}$$

Where VDW accounts for the Van der Waals interaction energy, calculated by considering all the atom pairs between the ligand and the protein; d_{ij} denotes the distance between the ligand atom i and the protein atom j; $d_{ij,0} = r_i + r_j$ the sum of Van der Waals radius of atom i and atom j. In X-score algorithm only heavy atoms contribute. Hydrogen atoms are neglected [Wang & Wang, 2001.].

Hydrogen bonding. Hydrogen bonding is the term that gives specificity in the bonding process. This interaction happens when two atoms get close enough and form a specific donor-acceptor pair. In the X-Score algorithm, *"a hydrogen bond donor is defined as a nitrogen or oxygen atom with a hydrogen attached; while an acceptor is defined as a nitrogen, oxygen, or fluorine atom with at least one valence electron to accept a hydrogen atom. All the atoms in the protein and the ligand are labeled as either donor (D), acceptor (A), donor/acceptor (DA), or none (N)"*, said the authors [Wang *et al.*, 1998].

A hydrogen bond has two parameters: the bond length, i.e the distance between D and A, and the bond angle, i.e., the angle between D-H...A.

Assuming that hydrogen bond has an ideal geometry and any deviation from it will weaken the strength of the hydrogen bond. The strength of a hydrogen bond is then computed by considering these three geometric descriptors [Wang & Wang, 2001.]:

$$HB_{ij} = f(d_{ij})f(\theta_{1,ij})f(\theta_{2,ij}) \tag{3.6}$$

The distance function $f(d)$ and the angular functions $f(\theta_1)$ and $f(\theta_2)$ in equation 3.6 are written in the following simple linear piece-wise forms;

$$f(d) = \begin{cases} 1.0 & \text{if} \quad d_0 \leq d_0 - 0.7\text{Å} \\ \frac{1}{0.7} \times (d_0 - d) & \text{if} \quad d_0 - 0.7\text{Å} < d < d_0 \\ 0.0 & \text{if} \quad d > d_0 \end{cases}$$

Where $d_0 = r_i + r_j$ is the Van der Waals distance between the donor and the acceptor and for the angular functions $f(\theta_1)$ and $f(\theta_2)$ we have respectively

$$f(\theta_1) = \begin{cases} 1.0 & \text{if} \quad \theta_1 \geq 120^\circ \\ \frac{1}{60} \times (\theta_1 - 60) & \text{if} \quad 60^\circ \leq \theta_1 < 120^\circ \\ 0.0 & \text{if} \quad \theta_1 \leq 60^\circ \end{cases}$$

$$f(\theta_2) = \begin{cases} 1.0 & \text{if} \quad \theta_2 \geq 120^\circ \\ \frac{1}{60} \times (\theta_2 - 60) & \text{if} \quad 60^\circ \leq \theta_2 < 120^\circ \\ 0.0 & \text{if} \quad \theta_2 \leq 60^\circ \end{cases}$$

The Hydrogen bonding interaction between the ligand and the protein is computed by summing up all the Hydrogen bonds:

$$HB = \sum_{i}^{ligand} \sum_{j}^{protein} HB_{ij} \tag{3.7}$$

It is important to emphasise that hydrogen bondings (HB) are stabilising interactions, and it is essentially a summation of the number of HB, so a large, positive HB is stabilising.

Deformation effect. According to wang :*"The deformation effect refers to the conformational changes that occur during the binding process. On one hand, this causes adverse entropic changes due to freezing of internal rotations of both the protein and its ligand and on the other hand, it can cause adverse enthalpic changes due to the strain energy exerted during binding. Based on the principles of statistical*

thermodynamics the entropic changes is usually estimated by using a constant value per rotatable bond that is frozen, but the enthalpic change is more difficult to elucidate" [Wang *et al.*, 1998].

X-Score algorithm uses the number of rotors in the ligand to estimate the term accounting for entropic and enthalpic change in the deformation effect. If a rotor is split into halves and assigned onto the two atoms involved then the term can also be written as [Wang *et al.*, 1998] :

$$RT = \sum_{i}^{ligand} RT_i \tag{3.8}$$

Where,

$$RT_i = \begin{cases} 1.0 & \text{if atom } i \text{ is involved in two rotors} \\ 0.5 & \text{if atom } i \text{ is involved in one or more than two rotors;} \\ 0.0 & \text{if atom } i \text{ is not involved in any rotor;} \end{cases}$$

We note that according to rotor-counting algorithm, RT_i should be 1.5 if atom i is involved in three rotors and 2.0 if four rotors are involved. This reduction in RT_i value reflects the consideration for offsetting the overstimation of conformational flexibility in the conventional algorithm. Although very crude, according to the authors, this reduction improves the accuracy X-Score [Wang & Wang, 2001.]

Here RT_i is the number of rotors in which ligand atom i is involved. A rotor is defined as acyclic $sp^3 - sp^3$ and $sp^3 - sp^2$ single bonds. Rotation of terminal $-CH_3$, $-NH_2$, or $-OH$, whose rotation does not produce any new conformation of heavy atoms, are note taken into account. The flexibility of cyclic portions of the ligand is ignored [Wang *et al.*, 1998].

It's worth to mention that attempt by the authors to incorporate the deformation effect of the protein into the computation did not improve the results, probably due to problems associated with estimating the magnitude and number of charges in amino side chain position on ligand binding.

Hydrophobic effect. Before protein and its ligand form a complex, both protein and ligand are solvated, and therefore a certain degree of desolvation takes place during binding that undergoes changes in entropy as well as in enthalpy. One of the consequences is that non-polar groups tend to favor each other, this is referred as "Hydrophobic effect". Accurate characterization is still very difficult to achieve as it involves complicated interactions such as ligand-water, protein-water, and water-water interactions before and after binding [Wang & Wang, 2001.].

It is important to note here however that this treatment of desolvation within X-Score is essentially an enthalpic one only: entropic component of desolvation are much more difficult to compute (not least because most water molecules are invisible in X-ray structure) and are essentially ignored by X-Score.

Many versions of algorithms have been proposed in other empirical scoring functions to compute hydrophobic terms. X-Score has implemented three classes of these algorithms.

1. **Hydrophobic surface algorithm.** The hydrophobic effect is assumed to be proportional to the buried hydrophobic surface of the ligand (Equation 3.9). Scoring function such as LUDI uses this algorithm [Bohm, 1994]. Since there are several types of molecular surfaces, X-Score choose to use the solvent-accessible surface (SAS). The radius of the solvent probe is set to 1.5 Å. The

surface areas of hydrogen atoms are attributed to their root atoms. Any part of the ligand surface is considered buried if it penetrates into the solvent-accessible surface of protein. Let also mention that only hydrophobic atoms are considered in Equation (3.9). The total amount of buried surface area is expressed in square Angstrom.

2. **Hydrophobic contact.** The hydrophobic effect is calculated by summing up the hydrophobic atom pairs formed between the ligand and the protein. In X-score it is calculated as in Equation (3.10).

3. **Hydrophobic matching algorithm.**
In the initial version of X-Score called SCORE [Wang *et al.*, 1998] this algorithm was adopted. According to this method, different parts of the ligand sense the protein differently because of the heterogeneous nature of the binding site. If a hydrophobic ligand atom is placed at a hydrophobic site of the protein, then it is expected to be favorable to the binding process [Wang & Wang, 2001.].

Hydrophobic surface is calculated as

$$HS = \sum_{i}^{ligand} SAS_i \tag{3.9}$$

And the Equation for Hydrophobic contact (pair) is given as follow:

$$HP = \sum_{i}^{ligand} \sum_{j}^{protein} f(d_{ij}) \tag{3.10}$$

Where

$$f(d) = \begin{cases} 1.0 & \text{if} \quad d \leq d_0 + 0.5\text{Å} \\ \frac{1}{1.5} \times (d_0 - 2.0 - d) & \text{if} \quad d_0 + 0.5\text{Å} < d \leq d_0 + 2.0\text{Å} \\ 0.0 & \text{if} \quad d > d_0 + 2.0\text{Å}. \end{cases}$$

Finally the overall hydrophobic matching (HM) between the ligand and the protein is computed as :

$$HM = \sum_{i}^{ligand} log\, P_i \times HM_i \tag{3.11}$$

Where HM_i is indicator function. It is set to 1 if hydrophobic atom i is placed in a hydrophobic environment; otherwise it is set to 0. $Log\, P_i$ refers to the hydrophobic scale of atom i, which is the contribution of atom i to the n-octanol/water partition coefficient ($Log\, P$) of the molecule. These scales play the role of weight factors to ensure that more hydrophobic atoms contribute more to the hydrophobic effect. The 'environment' of a given ligand atom is defined to consist of all the atoms on the protein which are within 6 Åfrom ligand atom. The hydrophobic of the environment is determined by summing up the hydrophobic scales of all its member atoms [Wang & Wang, 2001.].

Overall, X-Score expresses the binding affinity of a given protein-ligand complex in pK_d units calculated by summing up all terms described above. Since three different algorithms for modeling the hydrophobic effect have been implemented it result that X-Score consist of of three scoring functions:

$$
\begin{aligned}
pK_{d,1} &= C_{0,1} + C_{VDW,1} \times VDW + C_{H-bond,1} \times HB + C_{rotor,1} \times RT + C_{hydrophobic,1} \times HP, \\
pK_{d,2} &= C_{0,2} + C_{VDW,2} \times VDW + C_{H-bond,2} \times HB + C_{rotor,2} \times RT + C_{hydrophobic,2} \times HM \\
pK_{d,3} &= C_{0,3} + C_{VDW,1} \times VDW + C_{H-bond,3} \times HB + C_{rotor,3} \times RT + C_{hydrophobic,3} \times HS
\end{aligned}
\tag{3.12}
$$

Where $C_{0,i}$, $C_{VDW,i}$ and the coefficients $C_{H-bond,i}$, $C_{rotor,1}$, $C_{hydrophobic,i}$ $(i = 1, 2, 3)$ are adjustable parameters in the X-Score function: these default value coefficients are determined by regression analysis of the entire training set.

X-Score is therefore defined as a consensus scoring function which is the arithmetical average of the three equations of (3.12):

$$
X - Score = \frac{pK_{d,1} + pK_{d,2} + pK_{d,3}}{3} \tag{3.13}
$$

3.3 Regression analysis

A standard multivariate regression was carried out using the three equations in (3.12) on the training set to compute the weight or coefficient of each term. Table (3.2) lists all these coefficients.

Term	Coefficient
First equation in (3.12)	
VDW	0.004
H-Bond	0.053
Rotor	-0.061
Hydrophobic Pair	0.011
Regression constant	3.448
Second equation in (3.12)	
VDW	0.004
H-Bond	0.094
Rotor	-0.099
Hydrophobic Mathing	0.394
Regression constant	3.349
Third equation in (3.12)	
VDW	0.004
H-Bond	0.069
Rotor	-0.092
Hydrophobic surface	0.004
Regression constant	3.349

Table 3.1: Regression models of 3 equations in (3.12)

And the predictive coefficient of determination R^2 for the three program implemented in X-Score are respectively 0.318, 0.319 and 0.249 for the test data. The average value of these three scores gives the

coefficient of determination of X-Score which is 0.356. In term of goodness of fit we can say that only 35.6% of the test data are explained by the model proposed by X-Score. Investigating the statistical result for the training data set, R^2 for the respective equations in (3.12) is respectively 0.504, 0.546 and 0.571 yielding a R^2 of 0.591 for X-Score or in term of goodness of fit only 59.1% of training data set are explained by X-Score model which is quite a good proportion comparing to test data set where only 35%. The standard deviation for each algorithm in (3.12) is 1.51 for $HPSCORE$, 1.61 for $HMSCORE$ and 1.63 for $HSSCORE$ and taking the average of these values we get the deviation on the test data set for X-Score [Wang & Wang, 2001.].

A leave-one-out cross validation was performed and gave squared correlation coefficient (Q^2) of 0.480, 0.522 and 0.551 and a standard deviation (S_{PRESS}) of 1.62, 1.57 and 1.47.

3.4 Test data set and ranking

The value of any empirical scoring function resides in its capacity to reproduce a binding affinity that correlates well with experimental data. In the case of X-Score, 11 endothiapepsin complexes in the table below were used as the test data in the original X-Score work.

Applying X-Score to the test set [Wang *et al.*, 1998] they obtained a computed predictive correlation of 0.356 and a standard deviation of (s_{pred}) of 1.58 pK_d units (3.2 kJ/mol at 298 K).

PDB id	Resl (Å)	Protein/ligand	Exp	X-Score	Rank order Exp	Rank order Pred
1eed	2.0	endothiapepsin/PD-125754	4,90	6,15	11	11
1epo	2.0	endothiapepsin/CP-81282	7,96	8,84	3	2
1epp	1.9	endothiapepsin/PD-130693	7,16	6,58	6	10
2er0	3.0	endothiapepsin/L-364099	6,40	7,86	10	4
2er6	2.0	endothiapepsin/H-256	7,22	6,99	5	8
2er7	1.6	endothiapepsin/H-261	9.00	8,67	2	3
2er9	2.2	endothiapepsin/L-363564	7,80	7,83	4	5
3er3	2.0	endothiapepsin/CP-71362	7,10	6,9	7	9
4er1	2.0	endothiapepsin/PD-125967	6,62	7,69	9	6
4er2	2.0	endothiapepsin/pepstatin	9,30	9,27	1	1
4er4	2.1	endothiapepsin/H-142	6,80	7.00	8	7

Table 3.2: Test set and ranking

Figure (3.1 [Wang *et al.*, 1998] shows the correlation between the experimental and computed pK_d value of 11 endothiapepsin complexes in the test set. We notice however that X-Score is not a direct prediction of the experimental $-Log\, K_d$ as the fitted line has non-zero intercept (2.772) and the gradient is less than one (i.e 0.664). The test with 30 protein-ligand complexes provided a weak correlation as the regression yield 0.356 of R^2. Meaning that only $35,6\%$ of the data was explained by the regression line.

We also noted that from Table (3.2) whilst X-Score correctly predicts the strongest and weakest interactions in this small, focussed data set, it otherwise does not predict the correct rank order, which is potentially problematic in, for example, drug development applications.

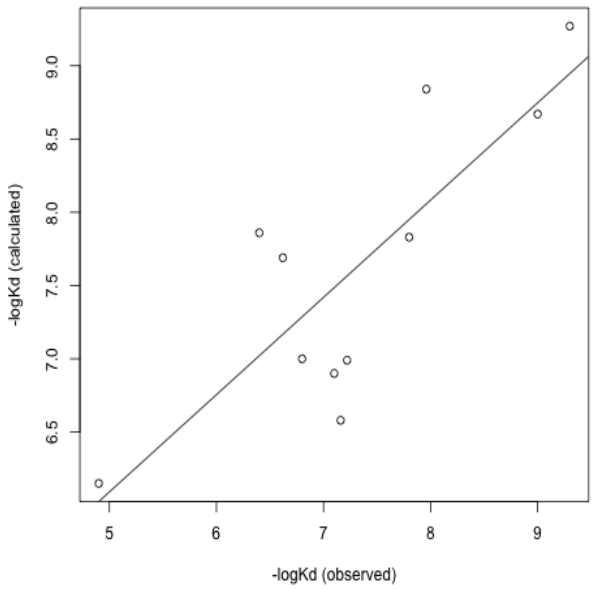

Figure 3.1: Correlation between the experimental and calculated pK_d values of 11 endothiapepsin complexes in the test set.

3.5 X-Score program

The final scoring function is written in C++ language. X-Score requires a protein file in PDB format together with the corresponding ligand file in MOL2 format as input to perform the computation. The program reads in the structure, assigns atom types to it and performs the calculation giving the negative logarithm of the dissociation constant for a given protein-ligand complex as output. The computation process for one complex is obtained within a second on SGI O2/R10000 workstation. The computational results are output into a text file in which the detailed information of each ligand atom, including the atomic binding score, is tabulated. Atomic binding scores are written into the Mol2 file which stores the ligand structure so that the user can display the using a visualization software such as SYBYL, VMD, etc [Wang & Wang, 2001.].

The authors of X-Score suggest that their approach perform well statistically compared to similar method. Indeed, as shown in Table 3.3, X-Score achieves the best regressional significance (F value) and smallest standard deviation in cross-validation (s_{PRESS}).

In Table 3.3 [Wang *et al.*, 1998] Samples is the number of complexes used in the training set. Terms indicate the number of terms in the scoring function, r^2 is the squared correlation given by regressional fitting, and s is the standard deviation in regressional fitting (kJ/mol), F is the Fisher significant ratio, q^2 is the squared correlation coefficient given by leave-one-out cross-validation and finally s_{PRESS} standard deviation in leave-one-out cross-validation (kJ/mol).

Approach	Bohm	Head	Gschwend	Eldridge	X-Score
Samples	45	51	103	82	170
Terms	5	13	8	5	6
r^2	0.762	0.85	0.745	0.710	0.777
s	7.9	5.8	7.2	8.0	6.6
F	32.1	17.8	39.6	-	57.8
q^2	0.696	0.78	0.701	0.658	0.743
s_{PRESS}	9.3	6.5	-	8.7	63

Table 3.3: Comparison of X-SCORE with other similar methods

3.6 Summary

In this chapter we have described X-Score, an empirical scoring function which can be used in structure-based drug design schemes. The parameters in the function aim to capture the essential energetics of the protein-ligand binding process. The model as designed is obtained after a regression of 200 complexes and evaluation against a test set.

Even Table 3.3 presents X-Score as producing a better binding affinity prediction compared to other scoring function, X-Score has however weaknesses among which we can cite:

1. entropic component of desolvation not considered
2. effectively double counts attractive dispersive force contributions (in both VDW and hydrophobic terms)
3. simplistic approach to Hydrogen bonding contributions (ignores, for example, the effect of charge on of strength of H-bonds, irrespective to distance and also appears to ignore sulphur atoms)
4. although overall regression-based coefficients give class-leading r^2 values etc. still gets rank order of ligand binding wrong within a given series (e.g. Table 3.2)

Therefore, there appears to be scope to globally re-parameterise X-Score for better application to individual protein-families (e.g protein kinases) or individual proteins (e.g Cell Division Kinase CDK2), as well as to explore the scope to locally reparameterise X-Score based on the dual effect of charged atomic pairwise interactions between protein and ligand on:

1. H-bond strength and
2. entropy of desolvation on ligand binding.

These therefore form the goals of the remaining chapters of this thesis.

4. Sensitivity and Re-parameterization of X-Score

In this chapter we assess the robustness of X-Score by measuring its sensitivity when changing successively the weight of each term and attempt to re-parameterize the coefficient tailored to a specific training set data in the group of human Kinase with emphasis on Epidermal growth Factor Receptor, the CDK2 (cell divsion protein kinase 2), MAPK (Mitogen-activated protein kinase) and the TK (Tyrosine Kinase).

It is worth mentioning here that we are mainly interested in dissociation constant (K_d) or inhibition constant (K_i) data and have focussed on these parameters as we aimed to optimise the X-Score coefficients for our data set of interest in this study. We began by using the data set provided by X-Score to recompute the X-Score's pK_d and to test for ourselves how well X-Score reproduce experimental pK_d.

Since we want to optimize parameters in X-score we are going to make constant use of linear regression which is the most famous and most widely used in optimization problems. The next two sections will consist of the statistical method used.

4.1 Statistical analysis

4.1.1 Linear Regression : simple linear regression

This section has been compiled from the book ([Draper & Smith, 1998]) and lecture notes from the South Africa National Bioinformatic (NBN) Course in 2011 at SANBI [Coetzer, 2011].

The simple linear regression model is a model in which we try to explain a endogenous variable Y as a function of a single exogenous variable x. It can be expressed as :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad 1 \leq i \leq n \tag{4.1}$$

where, $Y_1, Y_2, \ldots, Y_n$ are n endogenous observations variable (response variable); $x_1, x_2, \ldots, x_n$ are n exogenous (explicative variable); β_0 is the parameter associated to x-intercept; β_1 is the gradient;ϵ_i is an error term and n is the number of observations.

The variable $Y_1, Y_2, \ldots, Y_n$ are observations of one random variable, while $x_1, x_2, \ldots, x_n$ they are not random variable, they are known value (e.g through experiment). β_0 and β_1 are model parameters, β_0 is the x-intercept and β_1 is the slope of the line they are unknown and we must estimate them; $\epsilon_{1,2}, \ldots, \epsilon_n$ are error on one random variable.

Interpretation of β_0 and β_1

β_0 is interpreted like the average of value of Y when x is zero and β_1 is defined as the average increase of Y when x changes. If $\beta_1 = 0$, the distribution of Y in this case does not depend on x. If $\beta_0 = 0$, the model has no intercept.

The Figure 2.2 is an illustrative diagram of the simple linear regression model. The line represents the

average value of Y as a function of x. The observations values of Y is randomly distributed around the straight line representing this average. The error terms are differences between the observed values of Y and the fitted line. As the variance of these terms is constant in x, the average distance of the points to the regression line is the same for all values of x. Finally, the absence of pattern between the error terms means that the value of an error term is not influenced (linearly) by the value of other error terms.

Sometimes the average value of Y when x is zero is difficult to interpret. We can then choose a centered version of the model :

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4.2}$$
$$= \beta_0^* + \beta_1(x_i - \bar{x}) + \epsilon_i \tag{4.3}$$

In the centered version of the model, β_0 is interpreted as the average value of Y when x is $\bar{(}x)$.

Model parameters

In simple linear regression one estimate two parameters: β_0 and β_1. The variance of the residuals in the assumption of the normality were added to these two parameters. In this section, we show how to estimate these parameters. We will then focus on the distribution of β_0 and β_1 parameters and their interpretation.

In practice, one need to check the assumptions from estimate of the distribution of error terms, rather than from variables, itself. For this reason, it is possible to reformulate the hypothesis as follows:

$H_1 : \mathbb{E}(\epsilon_i) = 0$ for $i \leq i \leq n$
$H_2 : Var(\epsilon_i) = \sigma^2$ for $i \leq i \leq n$
$H_3 : Cov(\epsilon_i, \epsilon_j)$ for $i \leq i \leq n$

To these three assumptions, one need sometimes to add the assumption of normality of residuals as follows:

$H_4 : \epsilon_1, \epsilon_2, ..., \epsilon_n \sim N(0, \sigma^2)$

H_4 is generally a stronger assumption used because it allows to build confidence intervals and do hypothesis testing, depending on the size of data we are dealing with. If the size of data is big, we must make sure that hypothesis $H4$ is met otherwise we need to find an alternative test such as non-parametric test.

Estimation of β_0, β_1 and σ^2

Two approaches are generally accepted to estimate parameter of a statistical model: the least squares method and the maximum likelihood. This two approaches are equivalent under the assumption of normality.

textbfLeast Square Method

The aim here is to choose β_0 and β_1 to minimize the sum of squared residuals i.e :

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \Big(Y_i - \hat{Y_i} \Big)^2 = \sum_{i=1}^{n} \Big(Y_i - \beta_0 - \beta_i x_i \Big)^2 \tag{4.4}$$

where ϵ_i is an estimate of the error term, also called residual, Y_i is the observed value of the response variable and $\hat{Y_i}$ is the predicted value (i.e the position of Y_i if it was in the line). Since the function to be minimized has good properties (including smooth and convex), it can be minimized by taking the derivatives of the sum relative to β_0 and β_1, and then equating the derivative to zero and finally solving the system of two equations with two unknowns as follow:

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} & = 0 \\ \dfrac{\partial S}{\partial \beta_i} & = 0 \end{cases} \tag{4.5}$$

We do not get down to necessary mathematical simplification here to compute these estimators as it is shown in the next paragraph that the least squares method is equivalent to the method of maximum likelihood under the assumption of normality.

Maximum likelihood method

Assuming that the observations are independent, the likelihood is expressed as follows:

$$L\Big(\beta_0, \beta_1 \Big) = \prod_{i=1}^{n} f(Y_i) \tag{4.6}$$

The implementation of the maximum likelihood method requires the knowledge of the distribution of observations $Y_1, Y_2, ..., Y_n$. We have that $Y_i = (\beta_0 + \beta_1 x_i) + \epsilon_i$. Since the fourth assumption of regression (H_4: $\epsilon_{1,2}, ..., \epsilon_n \sim N(0, \sigma^2)$) and knowing that $(\beta_0 + \beta_1 x_i)$ is a constant, we have that $Y_i \sim N(\beta_0 + \beta_1 xi, \sigma^2)$. so

$$\begin{aligned} L\Big(\beta_0, \beta_1 \Big) &= \prod_{i=1}^{n} f(Y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2} \Big(\frac{Y_i - \beta_0 - \beta_1 x_i}{\sigma} \Big)^2 \Big) \end{aligned} \tag{4.7}$$

$\sum_{i=1}^{n} x_i$ It remain then to maximize the function $L\Big(\beta_0, \beta_1 \Big)$ or simply its logarithm $l\Big(\beta_0, \beta_1 \Big)$ where $l\Big(\beta_0, \beta_1 \Big)$ is defined as follow:

$$\begin{aligned} l\Big(\beta_0, \beta_1 \Big) &= \sum_{i=1}^{n} \{ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \Big(Y_i - \beta_0 - \beta_1 x_i \Big)^2 \} \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big(Y_i - \beta_0 - \beta_1 x_i \Big)^2 \end{aligned} \tag{4.8}$$

before further ado let's observe that in equation (4.8), we see that the first term contains no β_0 nor β_1. Therefore Maximize $l(\beta_0, \beta_1)$ is just to maximizing the second term. The latter being negative, maximize $L\Big(\beta_0, \beta_1\Big)$ is equivalent to minimize $\Big(Yi - \beta_0 - \beta_1 xi\Big)^2$, which corresponds exactly to the method of least squares.

The criteria (4.4) and (4.8) are equivalent. However, the maximum likelihood method is advantageous compared to the least squares method: it allows to estimate σ^2 directly. This allows us to find the distribution parameters and predictions and to test hypothesis in our linear model.

Estimate β_0

We First differentiate 4.8 with respect to β_0

$$\begin{aligned}\frac{\partial}{\partial\beta_0}\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(Y_i - \beta_0 - \beta_1 x_i\Big)\} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\Big(Y_i - \beta_0 - \beta_1 x_i\Big) \\ &= \frac{1}{\sigma^2}\Big(n\bar{Y} - n\beta_0 - n\beta_1\bar{x}\Big)\end{aligned} \tag{4.9}$$

we then set the derivative equal to zero

$$\begin{aligned} n\bar{Y} - n\beta_0 - n\beta_1\bar{x} &= 0 \\ \bar{Y} - \beta_0 - \beta_1\bar{x} &= 0 \\ \bar{Y} &= \beta_0 + \beta_1\bar{x}\end{aligned} \tag{4.10}$$

We observe that equation (4.10) means that the regression line necessarily passes through the point with coordinates $(\bar{x}, \bar{Y})$.

So that from (4.10) we deduce that :

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{x} \tag{4.11}$$

Estimate β_1

Differentiating equation (4.10)with respect to β_1, we get

$$\begin{aligned}\frac{\partial}{\partial\beta_1}\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(Y_i - \beta_0 - \beta_1 x_i\Big)^2\Big) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\Big(Y_i - \beta_0 - \beta_1 x_i\Big)x_i \\ &= \frac{1}{\sigma^2}\Big(\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n}\beta_0 x_- \sum_{i=1}^{n}\beta_1 x_i^2\Big).\end{aligned} \tag{4.12}$$

By setting this derivative equal to zero and substituting β_0 for the expression derived from equation

(4.10), we have

$$\begin{aligned}
\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} \beta_0 x_i - \sum_{i=1}^{n} \beta_1 x_i^2 = 0 \\
\sum_{i=1}^{n} x_i Y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \\
\sum_{i=1}^{n} x_i Y_i = \Big(\bar{Y} - \beta_1 \bar{x}\Big) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \\
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - \bar{Y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i}
\end{aligned} \tag{4.13}$$

It is then possible to re-write β_1 in the following equivalent form as:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\sum_{i=1}^{n} \Big(x_i - \bar{x}\Big)\Big(Y_i - \bar{Y}\Big)}{\sum_{i=1}^{n} \Big(x_i - \bar{x}\Big)^2} \\
&= \frac{\sum_{i=1}^{n} \Big(x_i - \bar{x}\Big) Y_i}{\sum_{i=1}^{n} \Big(x_i - \bar{x}\Big)^2} \\
&= \frac{S_{xy}}{S_{xx}}
\end{aligned} \tag{4.14}$$

Note that S_{xx} and S_{xy} are respectively called the corrected sum of squared of x and the corrected sum of the cross products of x and Y.

Estimate of σ^2

Finally, the last parameter of the model is to estimate is σ^2, the variance of the error terms. It is obtained in the same manner as β_1 and β_0 by differentiating the logarithm of the likelihood with respect to σ^2.

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2}\Big(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}\Big(Y_i - \beta_0 - \beta_1 x_i\Big)^2\Big) &= -\frac{n2\pi}{4\pi\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\Big(Y_i - \beta_0 - \beta_1 x_i\Big)^2 \\
&= \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\Big(Y_i - \beta_0 - \beta_1 x_i\Big)^2
\end{aligned} \tag{4.15}$$

By setting the derivative equal to zero and replacing the parameters β_0 and β_1 by their estimators, we have :

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon_i}^2 \quad ;\end{aligned} \tag{4.16}$$

where $\hat{\epsilon_i} = Y_i - \hat{Y_i}$ denote the residues.

4.1.2 Distribution for parameters

It should first be recalled that $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Since H_4 has been assumed that $\epsilon_1, ..., \epsilon_n \sim \mathcal{N}(0\,, \sigma^2)$ and $(\beta_0 + \beta_1 x_i)$ is a constant, it follows that $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i\,,\, \sigma^2)$. To find the law of the parameters, we take into account this facts.

Distribution for $\hat{\beta}_1$

Let's begin with the expression (4.13) for the estimator of β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.17}$$

Then setting the constant

$$a_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{4.18}$$

we can deduce that

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i \tag{4.19}$$

and under this form it is clear that $\hat{\beta}_1$ is a linear combination of Y_i and

$$\begin{aligned}\mathbb{E}[\hat{\beta}_1] &= \sum_{i=1}^{n} a_i \mathbb{E}[Y_i] \\ &= \sum_{i=1}^{n} a_i\left(\beta_0 + \beta_1 x_i\right) \\ &= \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i x_i \\ &= \beta_1\end{aligned} \tag{4.20}$$

It should be noted that for the last term of the equation, we used the following two equalities

$$\begin{aligned}
\sum_{i=1}^{n} a_i &= 0 \\
\sum_{i=1}^{n} a_i x_i &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{S_{xx}} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{S_{xx}} \\
&= \frac{S_{xx}}{S_x} \\
&= 1
\end{aligned} \tag{4.21}$$

We conclude that $\hat{\beta}_1$ is an unbiased estimator (or good estimator) of β_1

For the variance of $\hat{\beta}_1$

$$\begin{aligned}
\mathbb{V}ar[\beta_1] &= \mathbb{V}ar\left(\sum_{i=1}^{n} a_i Y_i\right) \\
&= \sum_{i=1}^{n} a_i^2 \mathbb{V}ar[Y_i] \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{S_{xx}^2}\mathbb{V}ar[Y_i] \\
&= \frac{S_{xx}}{S_{xx}^2}\mathbb{V}ar[Y_i] \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned} \tag{4.22}$$

we obtain the following distribution for $\hat{\beta}_1$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \tag{4.23}$$

Distribution for $\hat{\beta}_0$

$\hat{\beta}_0$ can be expressed as a linear combination of Y_i as follows:

$$\begin{aligned}
\hat{\beta}_0 &= \hat{Y} - \hat{\beta}_1 \bar{x} \\
&= \sum_{i=1}^{n} \frac{Y_i}{n} - \bar{x}\sum_{i=1}^{n} a_i Y_i \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - a_i \bar{x}\right) Y_i
\end{aligned} \tag{4.24}$$

Since a linear combination of normal random variables also follows a normal distribution,

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} & = 0 \\ \dfrac{\partial S}{\partial \beta_i} & = 0 \end{cases} \tag{4.25}$$

it follows that $\hat{\beta}_0$ is distributed according to a normal distribution. We need now to determine the parameters of this distribution, i.e the mean and variance values. We can calculate these quantities using the mean and variance of $\hat{\beta}_1$. for the mean we have that

$$\begin{aligned} \mathbb{E}[\hat{\beta}_0] &= \mathbb{E}[\bar{Y} - \hat{\beta}_1 \bar{x}] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i] - \bar{x}\mathbb{E}[\hat{\beta}_1] \\ &= \frac{1}{n}\sum_{i=1}^{n} \Big(\beta_0 + \beta_1 x_i\Big) - \bar{x}\beta_1 \\ &= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\ &= \beta_0 \end{aligned} \tag{4.26}$$

and the variance is calculated as follows

$$\begin{aligned} \mathbb{V}ar[\hat{\beta}_0] &= \mathbb{V}ar[\bar{Y} - \hat{\beta}_1 \bar{x}] \\ &= \mathbb{V}ar\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] + \bar{x}^2 \mathbb{V}ar[\hat{\beta}_1] - 2\bar{x}\mathbb{C}ov[Y, \hat{\beta}_1] \\ &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \end{aligned} \tag{4.27}$$

We obtain the following distribution for β_0:

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right) \tag{4.28}$$

Confidence intervals and hypothesis tests for β_0 and β_1

In order to have a confidence interval for β_0 and β_1 we will use the method where the distribution (equations (4.28) and (4.23)) of these parameters is required. For example, a confidence interval for β_1, one center and reduces the random variable to get

If σ^2 was known , a confidence level $(1-\alpha)\%$ is given by $[\beta_1 \pm z_{1-\alpha/2}\sigma/\sqrt{S_{xx}}]$. However, σ^2 is generally not known and must therefore be estimated. The maximum likelihood estimator of this

quantity is $\sum_{i=1}^{n} \hat{\varepsilon}_i$ which is biased. We therefore use rather unbiased estimator defined by $\sigma^2 = s^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i/(n-2)$. In this case, however, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{s^2/S_{xx}}}$ does not follow a normal distribution. The distribution is easily found.

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon_i}^2}{\sigma^2(n-2)}}} = \frac{U}{\sqrt{V/\nu}}$$

where $U = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}$ is distributed according to a standard normal distribution and $V = \frac{\sum_{i=1}^{n} \hat{\varepsilon_i}^2}{\sigma^2}$ follows a chi- square distribution with $\nu = (n-2)$ degrees of freedom (as the sum of the square of normal random variables) . We therefore obtain that

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon_i}^2}{\sigma^2(n-2)}}} \sim \frac{\mathcal{N}(0\,,\,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} \sim T_{n-2}$$

Since this is a ratio of a centered normal random variable and reduced the square root of a chi-squared variable divided by its degrees of freedom (n - 2), we obtain the distribution of T-Student with (n - 2) degrees of freedom (T_{n-2}). From the latter we derived directly the limits of a confidence interval at level of $(1-\alpha)\%$ for β_1 with the equation

$$\left[\hat{\beta}_1 \pm t_{n-2}(1-\alpha/2)\frac{s}{\sqrt{S_{xx}}}\right] \tag{4.29}$$

where $(1-\alpha/2)$ denotes the $(1-\alpha/2)$th quantile of a t-distribution with (n - 2) degrees of freedom. Based on the distribution found we can also implement hypothesis tests. The most often tested hypothesis concern the nullity of the slope (or null hypothesis) , i.e $\mathcal{H}_0$. $\beta_1 = 0$. Under $\mathcal{H}_0$, $\frac{\beta_1}{s/\sqrt{S_{xx}}} \sim t_{n-2}$ and $\mathcal{P}(|t_{n-2}| > |t_{obs}|)$, where t_{obs} denotes the calculated statistical test, gives us the p- value. If the p-value is smaller than the selected threshold α , we reject the null hypothesis that $\beta_1 = 0$. Conversely, if the p-value is greater than or equal the threshold, we can not reject the null hypothesis that $\beta_1 = 0$. in this case, we deduce that x has no explanatory power on Y. It is possible to build in the same way a confidence interval and hypothesis testing for β_0.

In the same fashion it is possible to build in a confidence interval and hypothesis testing for β_0. The confidence interval for the intercept can be expressed as :

$$\left[\hat{\beta}_0 \pm t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right] \tag{4.30}$$

In the case where the null hypothesis that the intercept $\beta_0 = 0$ can not be rejected, the model can be simplified as follows

$$Y_i = \beta_1 x_i + \varepsilon_i$$

4.1.3 Distribution of residues

In the model $Y_i = \hat{\beta}_0 + \hat{\beta}_1 xi$, Y_i denotes the *ith* predicted value, also called prediction, and $Y_i - \hat{Y}_i = \hat{\varepsilon}_i$ is called the *ith* residue.

The residuals of the model (, ...,) constitute part of the model (part of Y) unexplained by x. Residues can be seen as pseudo observations that tell us about the error terms.

Following the same approach as for the parameters, we calculate the distribution of residues by expressing them as a linear combination of Y_i as follows:

$$\begin{aligned}
\hat{\varepsilon_j} &= Y_j - \hat{j} \\
&= Y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j \\
&= Y_j - \sum_{i=1}^{n} (\frac{1}{n} - \alpha_i \bar{x}) Y_i - (\sum_{i=1}^{n} a_i Y_i) x_j \\
&= \sum_{i=1}^{n} d_i Y_i
\end{aligned} \tag{4.31}$$

where

$$a_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x}^2)} \quad \text{as defined above}$$

$$d_j = \begin{cases} -\frac{1}{n} + a_j(\bar{x} - x_j) & \text{if } i = j \\ \frac{1}{n} + a_i(\bar{x} - x_j) & \text{if } i \neq j \end{cases}$$

we have that

$$\begin{aligned}
\sum_{i=1}^{n} d_i &= \sum_{i=1}^{n} \Big(-\frac{1}{n} + a_i(x_i - x_j) \Big) \\
&= -1 + (\bar{x} - x_j) \sum_{i=1}^{n} a_i + 1 \\
&= 0
\end{aligned}$$

And

$$\begin{aligned}
\sum_{i=1}^{n} d_i^2 &= (1-\frac{1}{n})^2 + a_j^2(\bar{x}-x_j)^2 + 2(1-\frac{1}{n})a_j(\bar{x}-x_j) \\
&\quad + \sum_{i=1}^{n}\left((\frac{1}{n})^2 + a_i^2(\bar{x}-x_j)^2 - 2\frac{1}{n}a_i(\bar{x}-x_j)\right) \\
&= (\frac{n-1}{n})^2 + \frac{n-1}{n^2} + (\bar{x}-x_j)^2\sum_{i=1}^{n} a_i^2 + 2a_j(\bar{x}-x_j) \\
&\quad - \frac{2}{n}(\bar{x}-x_j)\sum_{i=1}^{n} a_i \\
&= \frac{(n-)^2+n-1}{n^2} + \frac{(\bar{x}-x_j)^2}{S_{xx}} + 2a_j(\bar{x}-x_j) \quad \text{car} \sum_{i=1}^{n} a_i = 0 \\
&= 1 - \left(\frac{1}{n} + \frac{(\bar{x}-x_j)^2}{S_{xx}}\right)
\end{aligned}$$

we deduce that

$$\begin{aligned}
\mathbb{E}[\hat{\varepsilon_j}] &= 0 \\
\mathbb{V}ar[\hat{\varepsilon_j}] &= \sum_{i=1}^{n} d_j^2 \mathbb{V}ar[Y_i] \\
&= \left[1 - \left\{\frac{1}{n} + \frac{(\bar{x}-x_j)^2}{S_{xx}}\right\}\right]\sigma^2 \\
&= (1-h_{jj})\sigma^2
\end{aligned} \tag{4.32}$$

Residues therefore follow a normal distribution

$$\mathcal{N}(1\,,\,(1-h_{jj})\sigma^2)$$

where $h_{jj} = \frac{1}{n} + \frac{(\bar{x}-x_j)^2}{S_{xx}}$

It should be noted that the sum of the residuals is always exactly equal to zero, which means that they are dependent variables.

4.1.4 Distribution of a prediction

It is common, although this is not always the case, the goal of regression is to predict a new value. We then distinguishes two cases:

1. Estimate Y_0, the mean value of Y predicted by the model for an exogenous variable value $x = x_0$.
2. Estimate the value any observation of Y when the exogenous variable x takes the value 0. This value will be denoted by $\hat{Y_0}^*$.

Prediction of the average value Y_0 for $x = x_0$

Let $\hat{Y_0}$ be a predictive value, the average value of Y predicted by the model when $x = x_0$. We have that $\hat{Y_0} = \hat{\beta_0} + \hat{\beta_1} x_0$. Again, we use the fact that $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i\,, \sigma^2)$ to find the distribution of $\hat{Y_0}$. Was first expressed $\hat{Y_0}$ as a linear combination of Y_i as follows:

$$\begin{aligned}
\hat{Y_0} &= \hat{\beta_0} + \hat{\beta_1} x_0 \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - a_i \bar{x} \right) Y_i + \left(\sum_{i=1}^{n} a_i Y_i \right) x_0 \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - a_i \bar{x} + a_i x_0 \right) Y_i \\
&= \sum_{i=1}^{n} c_i Y_i
\end{aligned} \tag{4.33}$$

where $c_i = \sum_{i=1}^{n} \left(\frac{1}{n} - a_i \bar{x} + a_i x_0 \right)$

We can show that $\sum_{i=1}^{n} c_i = 1$, that $\sum_{i=1}^{n} c_i x_i = x_0$ and that $\sum_{i=1}^{n} c_i^2 = \frac{1}{n} + \left\{ \frac{(\bar{x} - x_0)^2}{S_{xx}} \right\}$

We then deduce that

$$\mathbb{E}[\hat{Y_0}] = \sum_{i=1}^{n} c_i \mathbb{E}[Y_i] = \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 x_0 \tag{4.34}$$

$$\mathbb{V}ar[\hat{Y_0}] = \sum_{i=1}^{n} c_i^2 \mathbb{V}a[Y_i] = \left\{ \frac{1}{n} + \frac{(\hat{x} - x_0)^2}{S_{xx}} \right\} \sigma^2 \tag{4.35}$$

and that $\hat{Y_0} \sim \mathcal{N}(\beta_0 + \beta_1 x\,, \left\{ \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right\} \sigma^2)$

From this result, it is possible to calculate a confidence interval around the predicted mean value by replacing σ^2 by its estimator s^2. Using the same fashion as above, we obtain the confidence interval of level (1 -) to $\hat{Y_0}$ as follow:

$$\left[\hat{\beta_0} + \hat{\beta_1} x_0 \pm t_{n-2}(1 - \alpha/2) s \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}} \right] \tag{4.36}$$

Prediction of Y for $x = x_0$

It is important to understand that the confidence interval defined in equation (4.36 relates to the mean of $\hat{Y_0}$. To build what is called a prediction interval, we must consider both the variance related to the model (the estimation of parameters) and the variance of the error term (ε). However, equation (4.35) does not take into account the variance of the error term. The variance of this term is σ^2. If we denote the average prediction for $x = x_0$ by $\hat{Y_0} = \hat{\beta_0} + \hat{\beta_1} x_0$ and the prediction by $\hat{Y_0}^* = \hat{\beta_0} + \hat{\beta_1} x0 + \varepsilon$, we will have

$$\mathbb{V}ar[\hat{Y}_0] = \left\{ \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right\} \sigma^2$$

$$\begin{aligned} \mathbb{V}ar[\hat{Y}^*] &= \mathbb{V}ar[\hat{Y_0}^*] + \mathbb{V}ar[\varepsilon] \\ &= \left\{ \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right\} \sigma^2 + \sigma^2 \\ &= \left\{ 1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right\} \sigma^2 \end{aligned} \tag{4.37}$$

It is therefore possible to calculate the confidence interval at level (1 -) for $\hat{Y_0}^*$ and we is obtained

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2}(1 - \alpha/2) s \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}} \right] \tag{4.38}$$

The mean value of $\hat{Y_0}^*$ and of $\hat{}_0$ are equal because the mean value of the error terms is zero. It is important to note that the prediction interval is always wider than the confidence interval for $\hat{Y_0}$.

By analyzing the formulas (4.35) and (4.37), we find that the variances decrease when n increases or S_{xx} increases. This means that the higher the sample size is large and/or more there is variability in the values of the exogenous variable, the estimates will be more accurate. In contrary, more the point x_0 for which a prediction is required is far from the center ($\bar{x}$) of the explanatory variables more the prediction will be inaccurate. It is important to take these comments into consideration when planning experiments.

4.1.5 Analysis of Variance : ANOVA

The analysis of variance is one of the vast topic to linear modelling. Here, We will just give a little introduction needed to understand the regression. We will see in this subsection how the variability in the observations of the endogenous variable $Y_1, \ldots, Y_n$ can be decomposed as the linear regression model. This decomposition is the basis of the adjustment model tests. We have that

$$\begin{aligned} \sum_{i=1}^{n} (Y_i - \hat{Y})^2 &= \sum_{i=1}^{n} (Y_i - \bar{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + 2 \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(Y_i - \bar{Y}) \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2 \end{aligned}$$

If the sum of the total square $SS_{tot} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, the sum of squares due to regression $SS_reg = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ and the sum of residual squares $SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, then we have that

$$SS_{tot} = SS_{reg} + SS_{res} \tag{4.39}$$

We understand why this nomenclature for the various sums of squares by examining Table 2.3 that summarizes the analysis of variance. Equation (4.39) is predominant in regression. The total sum of squares SS_{tot} quantify the variability in Y_i. Equation (4.39) means that this variability is divided into two parts: the sum of squares due to regression (SS_{reg}) and the sum of square of the residual(SS_{res}).

The sum of squares due to regression SS_{reg} quantifies the variability in the estimates or prevision $\hat{Y}_i$, $i = 1, ..., n$. Since $\hat{Y}_i$, vary only with respect to x_i, $\quad i = 1, ..., n$, we have that this sum of square is the part of the variability in the Y_i explained by the fact that all observations do not have the same value for x_i.

The residual sum of squares SS_{res} measure the variability in the $(Yi - \hat{Y}_i)$. This variability is caused by the fact that the value of Y_i is not fully explained by the regression model.

The Table (4.1) presents the main elements of the analysis of variance.

Table 4.1: Table of analysis of variance in the case of simple linear regression having the source, the number of degrees of freedom, sum of squares, mean square and the Fisher statistic F

Source	Degree of freedom	Sum of squares (SS)	Mean Square (MS)	F
Model	1	$SS_{reg} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{x})^2$	$MS_{reg} = \frac{SS_{reg}}{1}$	MS_{reg}/MS_{res}
Residual error	n -2	$SS_{res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MS_{res} = \frac{SS_{res}}{n-2}$	
Total	n-1 = n-2 + 1	$SS_{tot} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$		

In Table (4.1) of Analysis of variance, each sum of the squares has a number of degrees of freedom of its own (shown in the second column of the table), namely:

- SS_{reg} has a single degree of freedom (df = 1) as in the simple linear regression model there is only one explanatory (exogenous) variable.
- In SS_{res} is associated to (n - 2) degrees of freedom since we estimated two parameters in the regression model (β_0 and β_1), so we have n observations minus 2 parameters.
- SS_{tot} has (n - 1) degrees of freedom because there are n observations minus one average. It is worth to note that the number of degrees of freedom of SS_{tot} is equal to the sum of the number of degrees of freedom of SS_{reg} and SS_res

The sum of squares due to regression (SS_{reg}) is the portion of variance explained by the model. The residual sum of squares (SS_{res}) is the part that the model does not explain. Intuitively, if the proportion of the variance explained by the model (SS_{reg}/SS_{tot}) is high, the model is good.

It is worth to emphasize that the residual mean square error (SM_{res}) is the estimate of the variance of the error terms. In fact

$$SM_{res} = \frac{MS_{res}}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}}{n-2} = s^2$$

.

The analysis of variance allows an overall test of the regression. We wish to test the null hypothesis $\mathcal{H}_0$: $Y_i = \beta_0 + \varepsilon_i$ against the alternative hypothesis $\mathcal{H}_1$: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The idea is to compare SS_{reg} to SS_{res} to assess whether exogenous variable explains a significant part of SS_{tot}. We have

already mentioned that if the proportion of variance explained by the model (SS_{reg}/SS_{tot}) is high, the model is good. To make the overall test, we instead use the ratio of mean squares (MS_{reg}/MS_res). To construct a test, it is first necessary to find the distribution of this ratio.

$$\begin{aligned}\frac{MS_{reg}}{MS_{res}} &= \frac{MS_{res}}{s^2} \\ &= \frac{S_{xy}^2/S_{xx}}{s^2} \\ &= \frac{\hat{\beta}_1^2 S_{xx}}{s^2} \\ &= \left(\frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}}\right)^2 \sim (\mathcal{T}_{n-2})^2 \sim \mathcal{F}_{1,n-2}\end{aligned}$$

In the last equality we used the fact that if the random variable T is distributed according to the $\mathcal{T}$ Student with k degrees of freedom (denoted $\mathcal{T}_k$), then the random variable T^2 follows the law of Fisher-Snedecor with 1 and k degrees of freedom (denoted $\mathcal{F}_{1,k}$).

So we will reject $\mathcal{H}_0$ at α level if $F \geqslant \mathcal{F}_{1,n-2}(1-\alpha)$. The p-value of the Fisher-Snedecor test indicates whether the exogenous variable has a significant effect on the value of the endogenous variable. A low p-value (e.g, if the p-value < 0.05) means that the effect of the exogenous variable is significant. As part of the simple linear regression, the Fisher-Snedecor test is exactly equal to the nullity of the slope $\mathcal{H}_0$: $\beta_1 = 0$. This case does not however generalize for the Multiple linear regression.

4.1.6 Quality of the model

To assess the quality of the model, we look at the following criteria:

- The Fisher test and the test of nullity of β_1. Both tests are equivalent under the simple linear regression. If the conclusion of the test is that $\beta_1 = 0$, the model is inadequate, that is to say that the explanatory variable (exogenous) does not have significant effect on the response variable (endogenous).
- Coefficient of determination (R^2).

$$R^2 = \mathbb{C}orr(Y\ ,\ \hat{Y}) = \left(\frac{\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y} - \bar{Y})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}}\right)^2 = \frac{SS_{reg}}{SS_{tot}}$$

When $R^2 = 0$, all the variability is due to random error and the model explains absolutely nothing about the value of Y_i. When $R^2 = 1$, all points are aligned on the regression line, that is to say that the model fit is perfect and the value of Y_i is an exact function of x_i. R^2 is interpreted as the percentage of variance explained by the model. A high value indicates that the model is good. The threshold for considering that R^2 is high, varies and quite subjective. In particular, it depends on the objectives of the regression and the domain of application.

4.1.7 Assumption in the model

After adjusting the model, it is important to check certain assumptions such as:

$\mathcal{H}_1$:	$\mathbb{E}[\varepsilon_i] = 0$	(linearity)
$\mathcal{H}_2$:	$\mathbb{V}ar[\varepsilon_i] = \sigma^2$	
$\mathcal{H}_3$:	$\mathbb{C}ov(\varepsilon_i, \varepsilon_j) = 0$	(No correlation)
$\mathcal{H}_4$:	ε_1, ..., ε_n	follow normal distribution(normality)

If these are not met, the linear regression is not an appropriate model and it will be necessary for example to transform variables or to use another type of model.

4.1.8 Multiple linear regression

Matrix notation for simple linear regression

The study of linear regression can be greatly deepened using matrix notation and results of linear algebra. Before addressing the multiple linear regression, we consider the matrix notation in the case of simple linear regression.

Let Y, the vector of dimension $(n \times 1)$ for endogenous variables, X a $(n \times 2)$ matrix of exogenous variables, β a (2×1) vector of the regression coefficients and the $(n \times 1)$ vector ε of error terms defined by

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

We have that

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Using these notations, the simple linear regression model defined in equation (4.57) can be written as:

$$Y = X\beta + \varepsilon \tag{4.40}$$

Introduction

Many regression problems involve several exogenous variables. Such approaches are called multiple regression models. Multiple linear regression remains one of the most applied statistical methods. Here we have more than one predictor.

The multiple linear regression make a relationship betwee one endogenous variable Y with multiple exogenous variables $(x_1, x_2, \ldots, x_p)$.

Model and Notation

The multiple linear regression model is a generalization of the simple linear regression model when considering several explanatory variables (exogenous). The equation of the multiple linear regression model can be written by the mathematical formula as follows:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \tag{4.41}$$

where,

- Y_i, $i = 1, \ldots, n$, represent the endogenous random variable.
- x_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, p$, denote the exogenous variables. These are known numbers, not random. It is possible to multiply β_0 by the variable $x_{i0} = 1$, $i = 1, \ldots, n$. In this case, β_0 represents a constant called Intercept.
- β_0 and β_j, $j = 1, \ldots, p$ denote the model parameters are unknown and therefore must be estimated.
- ε_i, $i = 1, \ldots, n$ are the unknow random variables error terms .

In its matrix notation, the model is written as follows :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or simply short hand notation

$$Y_{n\times 1} = X_{n\times p'} \beta_{p'\times 1} + \varepsilon_{n\times 1} \tag{4.42}$$

where,

- Y designates the vector containing the endogenous variable size $(n \times 1)$
- X notes the incidence matrix containing the exogenous variables (size $n \times p'$)
- β is the vector of regression coefficients (size $p' \times 1$)
- denotes of the error vector terms (size $n \times 1$)
- n denotes the number of observations
- p is the number of exogenous variables
- $p' = p + 1$ is the number of the exogenous variables plus intercept

Estimating regression parameters

As in the case of simple linear regression, we adopt again the Legendre's principle of least squares. One therefore seeks β vector which minimizes the sum of squared residuals ($\sum_{i=1}^{n} \hat{\varepsilon_i}^2$)

The residual vector can be expressed as

$$\hat{\varepsilon} = (\hat{\varepsilon_1}, \hat{\varepsilon_2}, \dots, \hat{\varepsilon_n})' = Y - X\hat{\beta}$$

$$\sum_{i=1}^{n} \hat{\varepsilon_i}^2 = \hat{\varepsilon}'\hat{\varepsilon} = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = f(\hat{\beta})$$

using matrices property we have that

$$\begin{aligned} f(\hat{\beta}) &= (Y' - (X'\hat{\beta})')(Y - X\hat{\beta}) \\ &= YY' - Y'X\hat{\beta} - (X\hat{\beta})' + (X\hat{\beta})'(X\hat{\beta}) \\ &= Y'Y - 2\hat{\beta}'X'Y - \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

In the last equality we used the fact that $Y'X\hat{\beta} = (X\hat{Y})'Y = \hat{\beta}'X'Y$.

We have then a quadratic function of $\hat{\beta}$. To minimize $\sum_{i=1}^{n} \hat{\varepsilon_i}^2$, we differentiate $f(\hat{\beta})$ with respect to $\hat{\beta}$ and we set that $f'(\hat{\beta}) = 0$ as follow

$$\frac{\partial f(\hat{\beta})}{\partial \hat{\beta}} = \begin{bmatrix} \frac{\partial f(\hat{\beta})}{\partial \hat{\beta}_0} \\ \frac{\partial f(\hat{\beta})}{\partial \hat{\beta}_1} \\ \vdots \\ \frac{\partial f(\hat{\beta})}{\partial \hat{\beta}_p} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

We get a $p+1$ system of equation with $p+1$ unknown for which it is possible to express in matrix form as follow:

$$\frac{\partial f(\hat{\beta})}{\hat{\beta}} = -2X'Y + 22X'X\hat{\beta} = 0$$

We need to have $X'X\beta = X'Y$. In the case where the matrix $X'X$ is invertible matrix we have that,

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{4.43}$$

If $X'X$ is not invertible matrix, one row of this matrix is a linear combination of others and it is necessary to get rid of this variable. In this case (4.43) has not unique solution.

4.1.9 Analysis of variance

In general, the n observed values $Y_1, \ldots, Y_n$, the endogenous variable are not all equal, that is to say that we observe the variability in the value of the endogenous variable. One goal of regression is to explain the largest possible part of the variability of the values from exogenous variables.

Thus, considering the decomposition of variability in the value of the following endogenous variable,

$$\begin{pmatrix} \text{variability of} \\ \\ Y_1, \ldots, Y_n \end{pmatrix} = \begin{pmatrix} \text{variability explained by} \\ \\ \text{the variability of} \, x_1, \ldots, x_n \end{pmatrix} + \begin{pmatrix} \text{unexplained variability} \\ \\ \text{(random fluctuation)} \end{pmatrix} \tag{4.44}$$

we want a model in which a large portion of the variability is explained by the variability in the exogenous variables. We can actually make the proposed (4.44) decomposition:

$$\begin{aligned} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}\right)^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + 2\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right) + \sum_{i=1}^{n} \left(\hat{Y} - \bar{Y}\right)^2 \\ &= \sum_{i=1}^{n} \left(\hat{Y} - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2 \end{aligned}$$

We can repeat the previous calculations in matrix form and define the following squares are:

$$\begin{array}{lll} SS_{tot} & = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 & = Y'Y - n\bar{Y}^2 \\ SS_{reg} & = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 & = \hat{Y}'\hat{Y} - n\bar{Y}^2 = \beta' X' X \hat{\beta} - n\bar{Y}^2 \\ SS_{res} & = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 & = \left(Y - \hat{Y}\right)' \left(Y - \bar{Y}\right) = \hat{\varepsilon}'\hat{\varepsilon} \end{array}$$

We can therefore express (4.44) in the following forms:

$$\begin{aligned} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 &= \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \\ Y'Y - n\bar{Y}^2 &= \left(\hat{\beta}' X' X \hat{\beta} - n\bar{Y}^2\right) + \hat{\varepsilon}'\hat{\varepsilon} \\ SS_{tot} &= SS_{reg} + SS_{res} \end{aligned} \tag{4.45}$$

For each square sum, a number of degrees of freedom is assigned. The degrees of freedom are in fact the number of independent terms we need to know the value in order to calculate the sum of squares. For example SS_{tot} has $n-1$ degrees of freedom, since only $n-1$ terms of $(Y_1 - \bar{Y}), \ldots, (Yn - \bar{Y})$ are independent (we know that their sum is 0, so if we known the value of $(n-1)$ of them, we can calculate the value of the nth).

Sums of squares and degrees of freedom are generally summarized in an analysis of variance Table (ANOVA). Table 4.2 summarizes the ANOVA standard Table, while Table 4.3 shows the ANOVA Table

decomposes the variability from 0 in three parts: one part due to intercept a second part due to external (exogenous) variable the third part due to random fluctuation. The F Column of the ANOVA Tables will be explained in the next section. In practice, we work almost exclusively with the ANOVA standard table .

Chart analysis of variance in the case of multiple linear regression with source, the number of degrees of freedom, sum of squares, mean square and Fisher statistic F

Table 4.2: Table of analysis of variance in the case of multiple linear regression with the source, the number of degrees of freedom, sum of squares, mean square and the Fisher statistic F

Source	Degree of freedom	Sum of squares (SS)	Mean Square (MS)	F
Model	p	$SS_{reg} = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$	$MS_{reg} = \frac{SS_{reg}}{p}$	MS_{reg}/MS_{res}
Residual error	$n - p'$	$SS_{res} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$	$MS_{res} = \frac{SS_{res}}{n-p'}$	
Total	$n - 1 = p + n - p'$	$SS_{tot} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$		

Table 4.3: Table of analysis of variance including the effect of β_0. Difference between total sum of squares and the total sum of squares corrected is that the first measures the variability of Y_i with respect to 0, while the second measures the variability of Y_i relative to their average.

Source	Degree of freedom	Sum of squares (SS)	Mean Square (MS)	F
β_0	1	$n\bar{Y}^2$	$n\bar{Y}^2$	$\frac{n\bar{Y}^2}{s^2}$
Model	p	$SS_{reg} = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$	$MS_{reg} = \frac{SS_{reg}}{p}$	MS_{reg}/MS_{res}
Residual error	$n - p'$	$SS_{res} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$	$MS_{res} = \frac{SS_{res}}{n-p'}$	
Total	$n - 1 = p + n - p'$	$SS_{tot} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$		

4.1.10 F-test of the overall significance of the regression

An important hypothesis testing in regression is to test if at least one exogenous variables explains a significant part of the variability in Y_i. This mean to test whether the data show some evidence against the null hypothesis $\mathcal{H}_0$: the exogenous variables explain nothing. Mathematically, an exogenous variable does not explain the value of Y_i if the corresponding regression coefficient is equal to 0. So we want to test

$$\begin{aligned} \mathcal{H}_0 &: \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\ \mathcal{H}_1 &: \text{at least one of the coefficients is not zero} \end{aligned}$$

Under $\mathcal{H}_0$, the regression model should not explain the variability in Y_i and therefore the ratio $\frac{SS_{reg}}{SS_{res}}$ should take a small value. By contrast under H_1, the regression model should explain some of the variability of Y_i and therefore $\frac{SS_{reg}}{SS_{res}}$ ratio should take a great value. To determine whether the value of the ratio is "small" or "large", we standardizes the ratio for the F statistic from the ANOVA table

$$F = \frac{SS_{reg}/p}{SS_{res}/(n-p')}) = \frac{SS_{res}/p}{s^2} = \frac{MS_{reg}}{MS_{res}}$$

Under H_0, the statistic $\mathcal{F}$ follows a Fisher-Snedecor with p degrees of freedom in the numerator and $(n-p')$ degrees of freedom in the denominator. We therefore reject H_0 at α level (that is, the data show that the model is not completely useless, or that there is relationship between the endogenous variable and at least one exogenous variables) when the F statistic is greater than or equal to the quantile $F_{p,n-p'}\,(1-\alpha)$

4.2 Recomputing X-Score

4.2.1 X-Score: Consensus scoring functions

The binary code, we received from the X-Score authors show that X-Score is a combination of three scoring functions as three algorithms are implemented from modeling the hydrophobic effect. The hydrophobic effect is calculated either by buried solvent-accessible molecular surface, or by the number of hydrophobic contacts between protein and the ligand or by the hydrophobic matching of the ligand with binding site. Recall that, the three functions in X-score are conceptually written as in equations (3.1), (3.2) and (3.3) in chapter 3.

The final X-score score is found by computing the mean of the three scores namely $HPSCORE$, $HMSCORE$ and $HSSCORE$ scores. i.e

$$X - Score = \frac{HPSCORE + HMSCORE + HSSCORE}{3} \tag{4.46}$$

4.2.2 Re-Computing X-Score

In this section we use the data set provided by X-Score (Table 4.4) to re-compute the X-score pK_d and find out for ourselves how well X-score performs.

4.2.3 Data set : X-Score provided data ("ad hoc")

The Table 4.4 gives both the predicted (Pred) X-Score pK_d and the observed or experimental values (Exp) of $-log\,K_d$ or $-log\,K_i$. The PDB ID for the protein-ligand complex used and the number of rotors in the ligand are also shown. The data set in Table 4.4 will be referring as "Ad hoc" data. These data was provided by the author of the program to test how well X-Score perform and from the data we computed the predicted score using X-Score.

Table 4.4: "ad hoc" data for all complex list used for X-score computation.

PDB ID	Exp	Pred	resolution in $\mathring{A}$	rotor	description
1a46	5.70	7.66	2.12	13	thrombin/beta-strand mimetic inhibitor
1a5g	10.15	7.38	2.06	14	thrombin/peptide inhibitor

Continued...

PDB	Exp	Pred	resolution	rotor	description
1abe	6.52	5.22	1.70	0	L-arabinose binding protein/L-arabinose
1abf	5.42	5.57	1.90	4	L-arabinose binding protein/D-fucose
1adb	7.69	7.67	2.40	11	alcohol dehydrogenase/CNAD
1add	5.74	5.82	2.40	2	adenosine deaminase/1-deaza-adenosine
1af2	3.10	5.40	2.30	2	cytidine deaminase/uridine
1apb	5.82	5.51	1.76	5	L-arabinose binding protein/D-fucose
1apt	9.40	6.94	1.80	20	penicillopepsin/pepstatin analogue
1apw	8.00	7.17	1.80	15	penicillopepsin/IvaValValDfo-N-methylamide
1b5g	8.00	7.45	2.07	11	serine protease/peptide mimetic inhibitor
1ba8	9.00	6.64	1.80	11	serine protease/peptide mimetic inhibitor
1bap	6.85	5.18	1.75	4	L-arabinose binding protein/L-arabinose
1bb0	8.36	6.75	2.10	11	serine protease/peptide mimetic inhibitor
1bbz	5.82	6.80	1.65	17	ABL tyrosine kinase/peptide ligand
1bcu	5.00	5.70	2.00	2	thrombin/proflavin
1bhf	4.38	5.85	1.80	20	tyrosine kinase P56LCK/ACE-IPA-GLU-GLU-ILE
1bra	1.82	5.10	2.20	1	trypsin mutant/benzamidine
1bxo	10.00	8.48	0.95	10	penicillopepsin/phosphonate inhibitor
1bzm	6.03	5.25	2.00	3	carbonic anhydrase I/sulfonamide drug
1cbx	6.35	5.90	2.00	5	carboxypeptidase A/L-benzylsuccinate
1cla	5.28	5.57	2.34	8	chloramphenicol acetyltransferase/chloramphenicol
1d3d	9.09	7.63	2.04	10	thrombin/benzo[B]thiophene inhibitor
1d3p	7.39	7.26	2.10	12	thrombin/benzo[B]thiophene inhibitor
1dhf	7.40	6.6	2.30	10	dihydrofolate reductase/folate
1dr1	5.57	5.48	2.20	5	dihydrofolate reductase/biopterin
1drf	7.44	6.89	2.00	10	dihydrofolate reductase/folate
1e96	5.22	6.92	2.40	8	RAC/P67phox
1ela	6.35	6.65	1.80	11	elastase/TFA-LYS-PRO-ISO
1etr	7.41	6.77	2.20	10	thrombin/MQPA
1ets	8.22	7.75	2.30	9	thrombin/NAPAP
1exw	3.90	6.04	2.40	15	palmitoyl protein thioesterase/hexadecylsulfonyl fluoride
1fkb	9.70	8.64	1.70	9	FK506 binding protein/rapamycin
1fkf	9.40	7.94	1.70	9	FK506 binding protein/FK506
1fmo	8.64	5.78	2.20	6	phosphotransferase/inhibitor PKI(5-24)
1hsl	7.30	5.08	1.89	4	histidine binding protein/histidine
1hvr	9.51	10.16	1.80	10	HIV-1 protease/XK263
1inc	8.00	6.35	1.94	8	porcine pancreatic elastase/benzoxazinone inhibitor
1mnc	9.00	6.56	2.10	9	neutrophil collagenase/hydroxamate
1ppc	6.16	6.91	1.80	10	trypsin/NAPAP
1pph	6.22	6.52	1.90	5	trypsin/3-TAPAP
1rbp	6.72	7.54	2.00	6	retinol binding protein/retinol
1rgk	4.31	5.14	1.87	7	ribonuclease T1/2'-AMP
1rgl	4.43	5.04	2.00	8	ribonuclease T1/2'-GMP
1rnt	5.18	5.45	1.90	7	ribonuclease T1/2'-GMP
1sre	4.00	6.76	1.78	4	streptavidin/HABA
1tet	6.20	4.34	2.30	6	IGG1 monoclonal fab fragment/CTP3
1tha	5.35	5.50	2.00	7	transthyretin/3 3'-diiodo-L-thyronine

Continued...

PDB	Exp	Pred	resolution	rotor	description
1tlp	7.56	7.45	2.30	11	thermolysin/phosphoramidon
1tmn	7.47	7.07	1.90	13	thermolysin/N-(1-carboxy-3-phenyl)-L-Leu-Trp
1tng	2.93	4.88	1.80	2	trypsin/aminomethylcyclohexane
1tnh	3.37	4.87	1.80	2	trypsin/4-fluorobenzylamine
1tni	1.70	4.95	1.90	5	trypsin/4-phenylbutylamine
1tnj	1.96	4.98	1.80	3	trypsin/2-phenylethylamine
1tnk	1.49	5.00	1.80	4	trypsin/3-phenylpropylamine
1tnl	1.88	5.17	1.90	2	trypsin/t-2-phenylcyclopropylamine
1yyy	5.09	5.77	2.10	9	serine protease/CVS1695
1zzz	5.13	5.59	1.90	9	serine protease/CVS1694
2ak3	3.86	5.73	1.90	7	adenylate kinase isoenzyme-3/AMP
2cgr	7.27	7.05	2.20	7	KAPPA fab fragment/antigen GAS
2csc	3.36	4.51	1.70	4	citrate synthase/D-malate
2ctc	3.89	5.53	1.40	4	carboxypeptidase A/L-phenyl lactate
2gbp	7.40	5.56	1.90	6	galactose binding protein/galactose
2pk4	4.32	4.37	2.25	6	plasminogen kringle 4/aminocaproic acid
2qwb	2.74	5.41	2.00	10	neuraminidase/sialic acid
2qwc	3.55	5.41	1.60	4	neuraminidase/neu5ac2en
2qwd	4.85	5.35	2.00	10	neuraminidase/4-amino-neu5ac2en
2qwe	7.48	5.70	2.00	10	neuraminidase/4-guanidino-neu5ac2en
2qwf	5.67	5.75	1.90	6	neuraminidase/ligand G20
2qwg	8.40	5.54	1.80	6	neuraminidase/ligand G28
2sns	6.70	5.69	1.50	4	staphylococcal nuclease/2'-deoxy-3' 5'-diphosphothymidine
2tmn	5.89	5.25	1.60	6	thermolysin/N-phosphory-L-leucinamide
2xim	2.28	4.55	2.30	4	D-xylose isomerase/xylitol
2xis	5.82	4.59	1.71	5	xylose isomerase/xylitol
3cla	4.94	4.31	1.75	8	chloramphenicol acetyltransferase/chloramphenicol
3cpa	4.00	5.60	2.00	5	carboxypeptidase A/glycyl-L-tyrosine
3fx2	9.30	7.49	1.90	10	flavodoxin/riboflavin monophosphate
3ptb	4.50	5.18	1.70	0	trypsin/benzamidine
3tmn	5.90	6.12	1.70	8	thermolysin/Val-Trp
4cla	5.47	5.52	2.00	8	chloramphenicol acetyltransferase/chloramphenicol
4sga	3.27	6.65	1.80	8	proteinase A/Ace-Pro-Ala-Pro-Phe
4tim	2.16	4.89	2.40	5	triosephosphate isomerase/2-phosphoglycerate
4tln	3.72	4.86	2.30	4	thermolysin/Leu-NHOH
4xia	1.54	4.81	2.30	5	D-xylose isomerase/D-sorbitol
5abp	6.64	5.49	1.80	6	L-arabinose binding protein/D-galactose
5cna	2.00	4.85	2.00	6	concanavalin A/a-Me-D-mannopyranoside
5p21	5.32	6.55	1.35	8	ras p21 protein/GPPNP
5sga	2.85	6.71	1.80	9	proteinase A/Ace-Pro-Ala-Pro-Tyr
5tln	6.37	5.93	2.30	8	thermolysin/benzylmalonyl-l-alanylglycine-p-nitroanilide
6abp	5.64	5.19	1.67	0	L-arabinose binding protein/L-arabinose
6rnt	2.37	5.25	1.80	7	ribonuclease T1/2'-AMP
6tim	3.21	4.81	2.20	6	triosephosphate isomerase/glycerol-3-phosphate
7abp	5.54	5.55	1.67	0	L-arabinose binding protein/D-fucose
7est	7.60	6.35	1.80	8	elastase/TFAP

Continued...

PDB	Exp	Pred	resolution	rotor	description
7tim	5.40	4.84	1.90	4	triosephosphate isomerase/phosphoglycolohydroxamate
7tln	2.47	5.24	2.30	7	thermolysin/CH2CO-Leu-OCH3
8abp	4.00	5.48	1.49	1	L-arabinose binding protein/D-galactose
8xia	2.95	4.46	1.90	4	D-xylose isomerase/D-xylose
9aat	8.22	5.55	2.20	6	aspartate aminotransferase/pyridoxal-5'-phosphate
9abp	8.00	5.43	1.97	6	L-arabinose binding protein/D-galactose

Here, Exp and Pred stand for experimental and predicted pK_d values respectively and we computed the pK_d values for these protein-ligand complexes using X-Score utility and open babel to prepare separately inputs files i.e protein and ligand files.

Model et postulats

The observed dependence of the X-Score pK_d value on the experimental value can be shown by a straight line. The regression line of variable X-score pK_d (or calculated score) (Y), on observed score (X) has the form i.e $\beta_0 + \beta_1 X$. Then the linear model can be written as :

$$\text{X-score}(Y) = \beta_0 + \beta_1 \text{ experimental score}(X)$$

The model assumptions include:

1. The linearity of the relationship
2. Homoscedasticity
3. Independence and
4. The normality of residuals

Since the objective is a predict the dissociation constant (pK_d), the assumption of normality must be satisfied.

Parameters estimation

The objective of the previous section was to show how linear regression work and how to apply it. There are a number of software packages available for performing least square regression. The regression parameters were estimated using **R** software [R Development Core Team, 2011.]. The results are summarized in Table 4.5. The Fisher test table analysis of variance shows that the model is useful ($p - value < 2.2e - 16$). The coefficient of determination R^2 is 0.6130677. The interpretation of this value is the following: about 61% of the variance of the computed dissociation constant is described by the experimental data. As expected, the adjusted R^2 is slightly less than R^2, but given the sample size, the difference is not very large. The coefficient of variation is 19

The coefficients β_0 and β_1 are both significantly different from zero; p-values are, respectively, $4.03e-16$ and $2e - 16$, which is well below the 5% threshold generally fixed. Confidence interval at 95% for $\hat{\beta}_0$ is constructed as follows:

$$\begin{aligned}
&\left[\hat{\beta}_0 \pm t_{n-2}(1-\alpha/2)s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\right] \\
&= \left[2.67207 \pm 1.98 \times \left(0.970275 \times \sqrt{\frac{(5.81222)^2}{488.1005}}\right)\right] \\
&= \left[2.67207 \pm 0.54104\right] \\
&= \left[2.1303\,,\, 3.2131\right]
\end{aligned}$$

The interval at 95% for $\hat{\beta}_1$ is

$$\begin{aligned}
&\left[\hat{\beta}_1 \pm t_{n-2}(1-\alpha/2)\sqrt{\frac{s^2}{S_{xx}}}\right] \\
&= \left[0.54446 \pm 1.98 \times \sqrt{\frac{0.9318271}{488.1005}}\right)\Bigg] \\
&= \left[0.54446 \pm 0.08651237\right] \\
&= \left[0.457947\,,\, 0.6309724\right]
\end{aligned}$$

Checking assumptions

Table 4.5: Analysis of Variance table

Sourcee	df	SS	$MS = SS/df$	F	$Pr > F$
Regression	1	144.71	144.712	155.29937	$< 2.2e-16$
Residual	97	91.31906	$s^2 = 0.9318271$		
Total	98	236.0313			

Root Mean Square Error (RMSE)	0.970275	R-squared	0.6131	Adjusted R-squared	0.6091

Coeffients	dof	Estimate value	Std.Error	t-value	Pr $>\mid t \mid$
Intercept	1	2.67207	0.27325	9.775	$4.03e-16$
Gradient	1	0.54446	0.04392	12.397	$< 2e-16$

The fitted equation is thus

$$\begin{aligned}
\widehat{Y} &= b_0 + b_1 X \\
&= 2.6721 + 0.5445X.
\end{aligned}$$

The foregoing form of $\widehat{Y}$ shows that $b_0 = 2.6721$. The fitted regression line is plotted in Figure (4.1). We can tabulate for each of 100 values of X_i for which Y_i observation is available, the fitted value $\widehat{Y}_i$,

and the residual $Y_i - \widehat{Y_i}$ as in Table (4.6). The residuals are given to the same number of places as the original data. They are the "estimates of the errors, ϵ_i" and can be written as $e_i = Y_i - \widehat{Y_i}$ in a corresponding notation.

We relied on "R" software to compute b_0 and b_1 the estimates value of β_0 and β_1 but we can check by a straight forward calculation that

$$b_0 = \bar{Y} - b_1\bar{X} \tag{4.47}$$

where $\bar{Y}$ and $\bar{X}$ are the mean of the observation of Y_i and X_i. Substituting equation (4.47) into equation (4.1) gives the estimated regression equation in the alternative form:

$$\widehat{Y} = \bar{Y} + b_1(X - \bar{X}) \tag{4.48}$$

From equation (4.48), we see clearly that if we set X $= \bar{X}$, then $\widehat{Y} = \bar{Y}$. This means that the center of gravity of the data set ($\bar{X}$, $\bar{Y}$) lies on the fitted line.

We now substitute equation (4.48) into the expression of "estimates errors" e_i written above

$$Y_i - \widehat{Y_i} = \Big(Y_i - \bar{Y}\Big) - b_1\Big(X_i - \bar{X}\Big)$$

,

We can sum both side to obtain,

$$\sum_{i=1}^{n}(Y_i - \widehat{Y_i}) = \sum_{i=1}^{n}(Y_i - \bar{Y}) - b_1\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

The above calculation suggests that the residuals sum to zero, in theory. In our case due to the rounding the residual sum is not exactly zero, but rather -0.028645 as it is in practice.

Table 4.6: Observations, Fitted values, and Residuals

$PDBID$	Y_i	$\widehat{Y_i}$	$Y_i - \widehat{Y_i}$
1a5g	7.66	8.20	-0.54
1a46	7.3	5.78	1.52
1abf	5.54	5.62	-0.08
1add	5.74	6.34	-0.60
1af2	5.35	4.36	0.99
1apb	5.52	5.84	-0.32
1apw	6.87	7.03	-0.16
1b5g	7.05	7.03	0.02
1ba8	6.07	7.57	-1.50
1bap	5.15	6.40	-1.25
1bb0	6.75	7.22	-0.47
1bbz	6.68	5.84	0.84
1bcu	6.25	4.46	1.79

Continued. . .

$PDB\,ID$	Y_i	$\widehat{Y}_i$	$Y_i - \widehat{Y}_i$
1bhf	5.83	5.06	0.77
1bra	5.02	3.66	1.36
1bxo	8.4	8.12	0.28
1cbx	5.86	6.13	-0.27
1d3d	7.42	7.62	-0.20
1d3p	6.73	6.70	0.03
1e96	6.57	5.70	0.87
1ela	6.29	6.14	0.15
1ets	7.47	7.15	0.32
1exw	5.94	4.80	1.14
1fkb	8.37	7.95	0.42
1fkf	7.78	7.79	-0.01
1fmo	5.72	7.38	-1.66
1hsl	4.97	6.59	-1.62
1hvr	9.96	7.85	2.11
1inc	6.26	7.76	-1.50
1ppc	6.81	6.03	0.78
1pph	6.48	5.90	0.58
1rbp	7.25	6.33	0.92
1rgk	5.18	5.02	0.16
1rgl	5.02	5.08	-0.06
1rnt	5.29	5.50	-0.21
1sre	6.45	4.77	1.68
1tet	4.42	5.98	-1.56
1tlp	7.38	6.78	0.60
1tmn	6.77	6.65	0.12
1tng	4.8	4.27	0.53
1tnh	4.79	4.51	0.28
1tni	4.74	4.85	-0.11
1tnj	4.88	3.74	1.14
1tnk	4.96	3.48	1.48
1tnl	5.05	3.70	1.35
1yyy	5.56	5.44	0.12
1zzz	5.53	5.47	0.06
2cgr	6.97	6.64	0.33
2ctc	5.2	4.79	0.41
2qwb	5.18	4.16	1.02
2qwc	5.22	4.61	0.61
2qwd	5.36	5.31	0.05
2qwe	5.45	6.74	-1.29
2qwf	5.46	5.76	-0.30
2tmn	5.24	5.88	-0.64
4sga	6.61	6.65	-0.04
4tim	4.88	3.85	1.03
5abp	5.45	6.29	-0.84
5tln	5.55	6.14	-0.59

Continued. . .

$PDB\,ID$	Y_i	$\widehat{Y_i}$	$Y_i - \widehat{Y_i}$
6abp	5.18	6.14	-0.96
6rnt	5.28	3.96	1.32
6tim	4.83	6.05	-1.22
7abp	5.51	6.19	-0.68
8abp	5.42	7.03	-1.61
9abp	5.43	7.03	-1.60
1bzm	6.03	5.96	0.07
1cla	5.28	5.55	-0.27
1dhf	7.4	6.70	0.70
1dr1	5.57	5.70	-0.13
1drf	7.44	6.72	0.72
1etr	7.41	6.71	0.70
1mnc	9	7.57	1.43
1tha	5.35	5.59	-0.24
2ak3	3.86	4.77	-0.91
2csc	3.36	4.50	-1.14
2gbp	7.4	6.70	0.70
2pk4	4.32	5.02	-0.70
2qwg	8.4	7.25	1.15
2sns	6.7	6.32	0.38
2xim	2.28	3.91	-1.63
2xis	5.82	5.84	-0.02
3cla	4.94	5.36	-0.42
3cpa	4	4.85	-0.85
3fx2	9.3	7.74	1.56
3ptb	4.5	5.12	-0.62
3tmn	5.9	5.88	0.02
4cla	5.47	5.65	-0.18
4tln	3.72	4.70	-0.98
4xia	1.54	3.51	-1.97
5cna	2	3.76	-1.76
5p21	5.32	5.57	-0.25
5sga	2.85	4.22	-1.37
7est	7.6	6.81	0.79
7tim	5.4	5.61	-0.21
7tln	2.47	4.02	-1.55
8xia	2.95	4.28	-1.33
9aat	8.22	7.15	1.07
1abe	6.52	6.22	0.30
1apt	9.4	7.79	1.61

4.3 Sensitivity of X-Score

We tested the sensitivity of X-Score on the target training data set of interest shown in Table 4.7 which is made of protein ligand complexes from the kinase protein family. Here I computed HPScore,

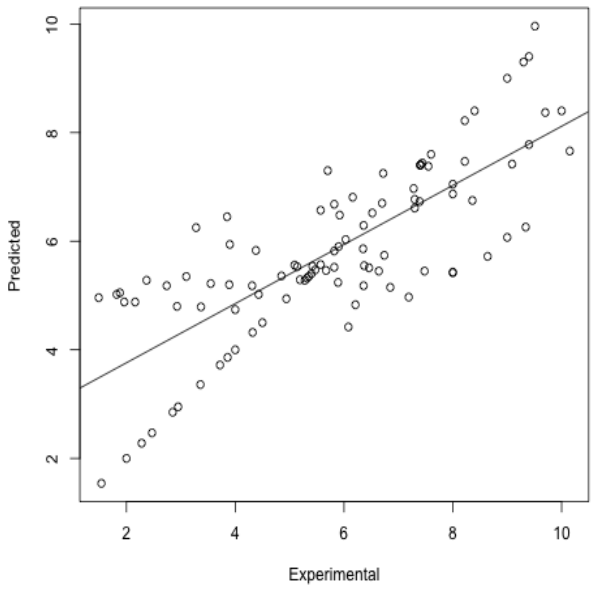

Figure 4.1: Plot of the data set and the least square line.

HMScore and HSScore values with X-Score utility and I prepared the protein and input file with open babel. Average gives pK_d values for X-Score as the finale score using the equation (4.46) which is a consensus between the three scoring functions.

Table 4.7: Kinase training data set

PDB /Ligand ID	HPScore	HMScore	HSScore	Average	Observed
1b38/ATP	6.02	6.19	5.81	6.01	6.60
1b39/ATP	5.93	6.07	5.70	5.90	6.92
1e1v/CMG	5.65	5.58	5.76	5.66	4.92
1e1x/NW1	5.46	5.39	5.51	5.46	5.89
1h1p/CMG	5.53	5.49	5.64	5.56	4.92
1h1s/4SP	6.56	6.77	6.59	6.64	8.22
1jsv/U55	5.48	5.68	5.23	5.46	5.70
1kv1/BMU	6.90	7.42	6.89	7.07	5.94
1kv2/B96	9.05	9.17	8.66	8.96	10.00
1pf8/SU9	6.12	6.11	5.51	5.91	7.51
1pxm/CK5	6.23	6.30	6.28	6.27	7.22
1pxn/CK6	6.30	6.32	6.18	6.27	7.15
1pxo/CK7	7.12	6.59	6.67	6.79	8.70
1pxp/CK8	6.33	6.51	6.37	6.41	6.66
1q4l/679	6.54	7.15	6.54	6.74	7.40
1y8o/ADP	5.21	5.30	4.99	5.17	5.89
2c6o/4SP	6.61	6.53	6.68	6.61	8.22

Continued...

PDB /Ligand ID	HPScore	HMScore	HSScore	Average	Observe
2clx/F18	5.54	5.76	5.28	5.53	4.88
2exm/ZIP	5.04	5.50	5.19	5.24	4.11
2fvd/LIA	6.51	6.65	6.23	6.46	8.52
2i6b/89I	6.39	6.79	5.65	6.28	7.17
2ity/IRE	6.07	6.22	5.79	6.03	7.27
2j6m/AEE	6.75	7.78	6.51	7.01	7.96
2pl0/STI	8.39	8.78	7.85	8.34	7.21
2qhm/7CS	5.93	6.00	5.62	5.85	6.18
2xnb/Y8L	10.30	10.67	9.22	10.06	6.83
2zkj/ADP	5.99	6.22	5.83	6.01	5.48
3d2r/ADP	5.92	6.14	5.76	5.94	5.48
3e64/5B3	6.74	7.61	6.64	7.00	7.11
3feg/AMP	5.04	5.10	4.85	5.00	4.35
3g0e/B49	6.75	6.83	6.44	6.67	7.70
3g0f/B49	6.53	6.68	6.24	6.48	7.66
3g15/HC6	7.04	7.43	6.14	6.87	7.00
3gfw/S22	6.80	7.74	6.52	7.02	7.57
3hec/STI	8.19	8.58	7.73	8.16	4.47
3heg/BAX	7.69	8.12	7.29	7.70	6.74
3huc/G97	7.37	7.63	6.65	7.22	5.99
3hv7/1AU	8.65	8.87	8.29	8.60	7.92
3jvs/AGY	6.71	7.45	6.58	6.91	6.54
3l8x/N4D	7.18	7.53	6.96	7.22	7.85
3lhj/LHJ	8.19	8.40	7.64	8.08	9.51
3new/3NE	7.08	7.16	6.32	6.85	5.00
3nga/3NG	7.38	7.72	6.59	7.23	8.00
3nyn/SGV	5.52	5.74	5.24	5.50	6.00

Table 4.8: Name of Kinase data set used in training set

PDB /Ligand ID	Name
1b38/ATP	Cell Division Protein Kinase 2/Adenosine Triphosphate
1b39/ATP	Cell Division Protein Kinase 2/Adenosine Triphosphate
1e1v/CMG	Cyclin Dependent Protein Kinase/Cyclo Hexyl methyl Guanine
1e1x/NW1	Cyclin Dependent Protein Kinase/Inhibitor NU6027
1h1p/CMG	Cell Division Protein Kinase 2/Inhibitor NU2058
1h1s/4SP	Cell Division Protein Kinase 2/Inhibitor NU206102
1jsv/U55	Cell Division Protein Kinase 2/Benzenesulfonamide
1kv1/BMU	p38 MAP kinase/Inhibitor 1
1kv2/B96	P38 MAP Kinase/BIRB796
1pf8/SU9	Cell Division Protein Kinase 2/SU9516
1pxm/CK5	Cell Division protein kinase 2/inhibitor 3
1pxn/CK6	Cell Division protein kinase 2/inhibitor 4
1pxo/CK7	Cell Division protein Kinase 2/Inhibitor
1pxp/CK8	Cell Division protein Kinase 2/Inhibitor
1q4l/679	Glycogen Synthase kinase-3 beta/Inhibitor I-5

Continued...

PDB /Ligand ID	Name
1y8o/ADP	Crystal structure of the PDK3-L2 complex/Adenosine Diphosphate
2c6o/4SP	Cell Division Protein kinase 2/Inhibitor
2clx/F18	Cell Division Protein kinase 2/CAN508
2exm/ZIP	Human CDK2 in complex with isopenteny ladenine
2fvd/LIA	Cyclin dependent kinase 2 (CDK2)/inhibitor
2i6b/89I	Adenosine kinase/Inhibitor
2ity/IRE	Epidermal Growth factor Receptor(EGFR)/IRESSA
2j6m/AEE	Epidermal Growth factor Receptor(EGFR)/AEE788
2pl0/STI	LCK bound to imatinib
2qhm/7CS	Crystal stucture of check I in complex with Inhibitor 2a
2xnb/Y8L	Crystal stucture of check I in complex with Inhibitor 2a
2zkj/ADP	Crystal stucture of Human PDK4-ADP complex
3d2r/ADP	Kinase Isozyme 4 in complex with ADP
3e64/5B3	'tyrosine-proteine kinase Jack2 /Inhibitors
3feg/AMP	Choline/Ethanolamine Kinase/Adenosine Monophosphate
3g0e/B49	Mast/Setm cell growth factor receptor/Sunitinib
3g0f/B49	Mast/Setm cell growth factor receptor/Sunitinib
3g15/HC6	Choline kinase alpha/Hemicholinium
3gfw/S22	Dual specificity protein kinase TTK/Pyrolo-pyridin ligand
3hec/STI	P38 in complex with Imatinib
3heg/BAX	P38 in complex with Sorafenib
3huc/G97	Human P38 Kinase in complex with RL40
3hv7/1AU	Human P38 kinase in complex with RL38
3jvs/AGY	Serine/Threnonine-protein kinase chk1
3l8x/N4D	Mitogen activated protein kinase 14/inhibitor
3lhj/LHJ	Mitogen activated protein kinase 14/pyrazolo pyridinone Inhibitor
3new/3NE	P38-alpha complexed with compound 10
3nga/3NG	Casein Kinase II subunit alpha in complexe with Cx-4945
3nyn/SGV	G protein-couple receptor kinase 6 in complex sangivamycin

To do this we assigned a weighting to individual terms and following this, we re-compute X-Score and test the output against real experimental data to examine the robustness of the weighting used.

First we repeated the same calculation as we did in above here and Table (4.1) gives the analysis of variance summary.

Table 4.9: Analysis of variance for the kinase data set

Source of variation	Degrees of Freedom (df)	Sum of Squares (SS)	MS_{reg}
Due to regression	1	12.19212	12.19212
About Regression	42	37.07508	$s^2 = 0.862211$
Total, corrected	43	49.2675	-

And R^2 statistics is given by

$$R^2 = \frac{12.19212}{49.2675} = 0.2474678$$

The fitted regression equation for this case is then given by the equation $\hat{Y} = 2.463591 + 0.643372X$ which explain only 24.74% of the total variation in the kinase data set about the average $\bar{Y}$. In other word the proportion of data which is not explained by the fitted regression line equation is 75.25%. This is a large proportion of data.

4.3.1 Standard Deviation of the Gradient : Confidence interval for β_1

Before testing the sensitivity of X-Score, we computed the standard deviation of the gradient of our fitted regression line.

From equation (4.13) we know that

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \tag{4.49}$$

Using this convention to write the numerator of b_1 in symmetrical form by letting the term $\sum(X_i - \bar{X})\bar{Y} = \bar{Y}\sum(X_i - \bar{X}) = 0$, the gradient b_1 becomes

$$b_1 = \frac{\sum(X_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \tag{4.50}$$

And the variance of the function (b_1 is

$$Var(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = \frac{\sigma^2}{S_{XX}} \tag{4.51}$$

The standard deviation of b_1 is the square root of the variance, that is

$$sd(b_1) = \frac{\sigma}{\left\{\sum(X_i - \bar{X})^2\right\}^{1/2}} = \frac{\sigma}{S_{XX}^{1/2}} \tag{4.52}$$

If σ is unknown and we use the estimates in its place, assuming the model is correct, the estimated standard deviation of the gradient b_1 is given by

$$est.sd(b_1) = \frac{s}{\left\{\sum(X_i - \bar{X})^2\right\}^{1/2}} = \frac{s}{S_{XX}^2} \tag{4.53}$$

The estimated standard deviation is the standard error, se.

Confidence interval for β_1 We assign $100(1-\alpha)\%$ confidence limits for β_1 by calculating

$$b_1 \pm \frac{t(n-2, 1-\frac{1}{2}\alpha)s}{\left\{\sum(X_i - \bar{X})^2\right\}^{1/2}} \tag{4.54}$$

Where $t(n-2, 1-\frac{1}{2}\alpha)$ is the $100(1-\alpha)$ percentage point of a t-distribution, with $(n-2)$ degrees of freedom.

From computation in our case it follow that

$$\begin{aligned} Var(b_1) &= \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \\ &= \frac{\sigma^2}{78.85916} \end{aligned}$$

$$\begin{aligned} est.Var(b_1) &= \frac{s^2}{78.85916} \\ &= \frac{0.862211}{78.85916} \\ &= 0.01093355 \\ se(b_1) &= \sqrt{est.Var(b_1)} = 0.1045636 \end{aligned}$$

Let $\alpha = 0.05$, so that $t(42, 0.975) = 2.0189$. Then the 95% confidence limits for β_1 are

$$b_1 \pm \frac{t(42, 0.975)s}{\left\{\sum(X_i - \bar{X})^2\right\}^{1/2}}$$

or

$$0.643372 \pm 2.463591 \times 0.172826$$

providing the interval

$$0.217599 \leqslant \beta_1 \leqslant 1.069145$$

The true value of β_1 lies in interval $(0.217599\, to\, 1.069145)$, and this calculation is made with 95% confidence. Since the goodness of fit decreases as the gradient increases we restricted our confidence interval to $(0.217599\, to\, 1)$

Test of Null Hypothesis We tested the null hypothesis that the true β_1 is zero, or that there is no straight line sloping relationship between computed score (Y) and experimental score (X).

$$H_0 : \beta_1 = 0 \qquad\qquad H_1 : \beta_1 \neq 0 \tag{4.55}$$

t value is given by

$$t = \frac{r(n-2)}{\sqrt{1-r^2}} \tag{4.56}$$

where r is the r-value or the correlation coefficient (computed here as the square root of the coefficient of determination) and n is the size of our data.

$$r = \sqrt{R^2} = \sqrt{0.2474678} = 0.4974614$$

and evaluate

$$t = \frac{r(n-2)}{\sqrt{1-r^2}} = 3.716453$$

since $\mid t \mid = 3.716453$ is greater that the critical value of $t(42\,,\,0.975) = 2.0189$, $H_0 : \beta = 0$ is rejected. We therefore reject the idea that a linear relationship between the computed (Y) values and the observed values does not exist.

Sensitivity of the Hydrogen Bonding term once we had compute the standard deviation of the gradient β_1 and the confidence interval where the true value of β_1 lies we were now ready to investigate the effect of assigning different weighting to the individual terms on X-Score performance. Following this, we re-computed the X-Score pK_d and tested the output against experimental data to examine the robustness of weighting used. We will begin by adjusting the weight of the hydrogen bonding term. We know that the true value of the gradient β_1 lies in the interval

$$0.217599 \leqslant \beta_1 \leqslant 1$$

During the calculation, we dropped the value of the gradient that is outside of the limits interval and for those in the interval we considered only those for which the $p - value \leq 0.025$. Table 4.9 gives all the Gradient, intercept, r-value, P-value and the standard error when the coefficient of the hydrogen bonding term undergoes small perturbation in equations ((3.1), (3.2), (3.3). The coefficient were generated by letting the weight of each term float with a small change of 0.001.

By graphing the data from the Table 4.11 we have the figure below.

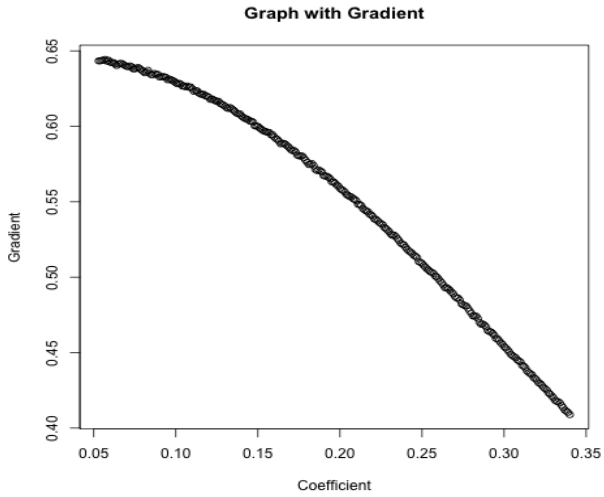

Figure 4.2: Shows a plot of the gradient vs the coefficient for the data in Table 4.11

The optimal value of the gradient is reached when the coefficients are respectively $CHB1 = 0.053$, $CHB1 = 0.094$, $CHB1 = 0.069$. These are default value from X-Score which also yield the optimal value of the r-value (r-value is the square root of R^2). The r-value is then equal to 0.498092 or using R^2 statistics, 0.248095. Table 4.11 is organized in descending order with respect to the gradient. We observe in the graph and Table 4.11 that the gradients decrease when the coefficients increase.

We also show in Figure 4.3, the graph of gradient vs coefficient for both increased and decreased coefficients. We noticed that even when we decrease the coefficient the gradient did not improve much but a slight improvement of the R^2 statistics is achieved.

The table below gives the optimal value reach by the gradient and $r - value$ while we decrease the coefficient of the hydrogen bonding term.

Table 4.10: Linear regression : statistics summary 1

HBP	HBM	HBS	Gradient	intercept	r-value	p-value	std_err
-0.117000	-0.076000	-0.101000	0.609975	2.894121	0.523389	0.000267	0.153232
0.009000	0.050000	0.025000	0.648616	2.484863	0.510719	0.000397	0.168481

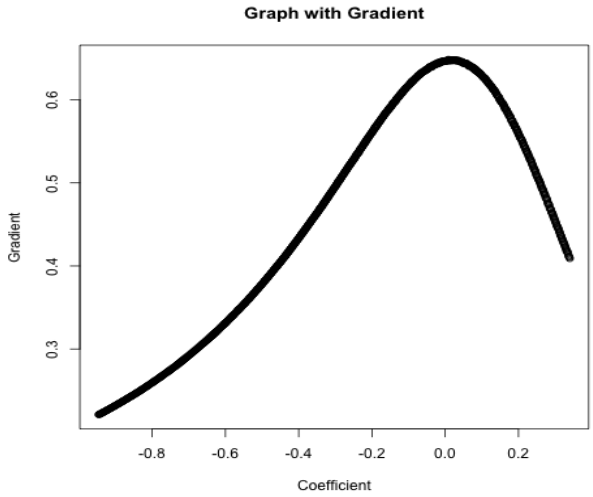

Figure 4.3: Plot of gradient vs coefficient for the hydrogen bonding.

X-Score sensitivity to change in the weighting of Van der Waal Term In the same way we then tested the sensitivity of X-Score to the perturbation of the Van der Waal term coefficient. Doing as above we obtain the gradient, intercept and r-value as in Table 4.11. The highest optimal value of the gradient is attained when the coefficients decrease down to $HBP = HBM = HBS = 0.002$ in the Van der Waal term coefficient in equations 3.1, (3.2) and (3.3) for which the gradient is 0.923038. The highest score of R^2 is reached where the gradient is at its optimal value which is 0.261734.

Table 4.11: Sensitivity of gradient to the Van der Waal term

HBP	HBM	HBS	$Gradient$	$intercept$	$r-value$	$p-value$	$std-err$
0.004000	0.004000	0.004000	0.643372	2.463591	0.498092	0.000580	0.172826
0.005000	0.005000	0.005000	0.551103	2.750216	0.490889	0.000716	0.150923
0.006000	0.006000	0.006000	0.479989	2.980592	0.484305	0.000865	0.133797
0.007000	0.007000	0.007000	0.422544	3.180165	0.477422	0.001049	0.119998
0.008000	0.008000	0.008000	0.377815	3.333343	0.472623	0.001197	0.108704
0.009000	0.009000	0.009000	0.340523	3.468087	0.467705	0.001368	0.099299
0.010000	0.010000	0.010000	0.309522	3.582772	0.463088	0.001548	0.091409
0.011000	0.011000	0.011000	0.283596	3.679462	0.459544	0.001701	0.084574
0.012000	0.012000	0.012000	0.261556	3.762433	0.456139	0.001859	0.078739
0.013000	0.013000	0.013000	0.242800	3.832550	0.453452	0.001993	0.073639
0.014000	0.014000	0.014000	0.226214	3.897293	0.450702	0.002139	0.069135
0.003000	0.003000	0.003000	0.766708	2.098764	0.506702	0.000449	0.201289
0.002000	0.002000	0.002000	0.923038	1.701471	0.511600	0.000386	0.239206

Figure 4.4 shows a plot of the gradient vs the coefficient

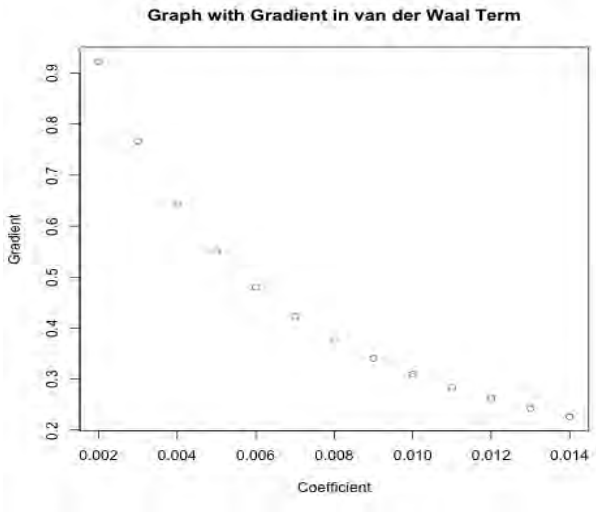

Figure 4.4: Plot of gradient vs coefficients with van der Waal Term

X-Score's sensitivity to changes in the weighting of the hydrophobic terms The results obtained when we perturbed the coefficient of the hydrophobic terms are shown in Table 4.12.

From Table 4.12 we see that the highest gradient value of $0,65$ is reached when we use the default coefficients ($HP = 0,011\, HM = 0,395$ and $HS = 0,004$). But interestingly the correlation coefficient attained the optimal value of 0.553012 (or $R^2 = 30.58\%$) when the coefficients were $HP = 0,42\, HM = 0,425$ and $HS = 0,035$ compared to r-value of $0,497543$ (or $R^2 = 24.75\%$) for the default value. This suggest that when the gradient decrease from 0.650051 to 0.305819 the proportion of data explained by the fitting line increases by approximately 5%.

Table 4.12: Sensitivity of the gradient after perturbation of the Hydrophobic term coefficient

HP	HM	HS	Gradient	Intercept	r-value	p-value	std-err
0,042	0,425	0,035	0,305819	3,945552	0,553012	0,000099	0,071095
0,037	0,42	0,03	0,339931	3,767189	0,552982	0,000099	0,079032
0,04	0,423	0,033	0,318621	3,878413	0,552921	0,000099	0,074089
0,039	0,422	0,032	0,325497	3,842707	0,552899	0,000099	0,075692
0,035	0,418	0,028	0,355348	3,688089	0,552809	0,0001	0,082653
0,038	0,421	0,031	0,33233	3,807612	0,552795	0,0001	0,077302
0,036	0,419	0,029	0,347415	3,729346	0,552776	0,0001	0,080815
0,041	0,424	0,034	0,311976	3,913924	0,552748	0,0001	0,072576
0,043	0,426	0,036	0,299668	3,978193	0,552712	0,0001	0,06972
0,044	0,427	0,037	0,293866	4,008486	0,552637	0,0001	0,068383
0,034	0,417	0,027	0,363649	3,645759	0,552592	0,0001	0,084632
0,033	0,416	0,026	0,37228	3,601466	0,55246	0,000101	0,08667
0,045	0,428	0,038	0,288206	4,038694	0,55246	0,000101	0,067097
0,047	0,43	0,04	0,277519	4,095405	0,55232	0,000101	0,064633
0,032	0,415	0,025	0,381169	3,556653	0,552125	0,000102	0,088817

Continued...

HP	HM	HS	Gradient	Intercept	r-value	p-value	std-err
0,046	0,429	0,039	0,282577	4,069259	0,552118	0,000102	0,065845
0,031	0,414	0,024	0,390707	3,507406	0,55205	0,000102	0,091058
0,049	0,432	0,042	0,267459	4,149049	0,551922	0,000103	0,062354
0,048	0,431	0,041	0,27222	4,124602	0,551754	0,000103	0,063492
0,05	0,433	0,043	0,262586	4,175698	0,551615	0,000104	0,061267
0,051	0,434	0,044	0,258057	4,199578	0,551598	0,000104	0,060213
0,03	0,413	0,023	0,400097	3,461381	0,551542	0,000104	0,093369
0,052	0,435	0,045	0,253569	4,224007	0,551308	0,000105	0,059211
0,053	0,436	0,046	0,249255	4,247345	0,551212	0,000105	0,058218
0,054	0,437	0,047	0,245054	4,270113	0,550953	0,000106	0,057275
0,029	0,412	0,022	0,410026	3,412269	0,550889	0,000107	0,09585
0,055	0,438	0,048	0,241078	4,291505	0,550749	0,000107	0,056376
0,056	0,439	0,049	0,237188	4,312262	0,550681	0,000107	0,055476
0,028	0,411	0,021	0,42084	3,357563	0,550525	0,000108	0,098471
0,057	0,44	0,05	0,233335	4,333488	0,550438	0,000108	0,05461
0,059	0,442	0,052	0,22616	4,371841	0,550253	0,000109	0,052956
0,058	0,441	0,051	0,229625	4,353613	0,550207	0,000109	0,053774
0,06	0,443	0,053	0,222514	4,392677	0,549717	0,000111	0,052175
0,027	0,41	0,02	0,431393	3,306921	0,549562	0,000112	0,101194
0,061	0,444	0,054	0,219137	4,411094	0,549546	0,000112	0,051406
0,026	0,409	0,019	0,443153	3,24817	0,549344	0,000113	0,104011
0,025	0,408	0,018	0,454293	3,19602	0,547705	0,000119	0,107083
0,024	0,407	0,017	0,466706	3,135936	0,546782	0,000123	0,110274
0,023	0,406	0,016	0,479538	3,074736	0,545307	0,000129	0,113743
0,022	0,405	0,015	0,492764	3,012612	0,544011	0,000135	0,117276
0,021	0,404	0,014	0,506347	2,949591	0,542229	0,000144	0,121071
0,02	0,403	0,013	0,520356	2,886161	0,540359	0,000153	0,12503
0,019	0,402	0,012	0,534194	2,825932	0,537369	0,000169	0,129362
0,018	0,401	0,011	0,549793	2,755425	0,535239	0,000181	0,133884
0,017	0,4	0,01	0,564366	2,694261	0,531516	0,000205	0,13878
0,016	0,399	0,009	0,579132	2,634062	0,527322	0,000235	0,143988
0,015	0,398	0,008	0,595069	2,568226	0,523641	0,000265	0,149389
0,014	0,397	0,007	0,609569	2,515451	0,518358	0,000313	0,155173
0,013	0,396	0,006	0,623715	2,4665	0,512637	0,000374	0,161193
0,012	0,395	0,005	0,637446	2,423966	0,505592	0,000464	0,167847
0,011	0,394	0,004	0,650051	2,390256	0,497543	0,00059	0,174876

X-Score's sensitivity to change in the weighting of the rotor term The result obtained when we perturbed the coefficient of the rotor term is shown in Figure 4.6.

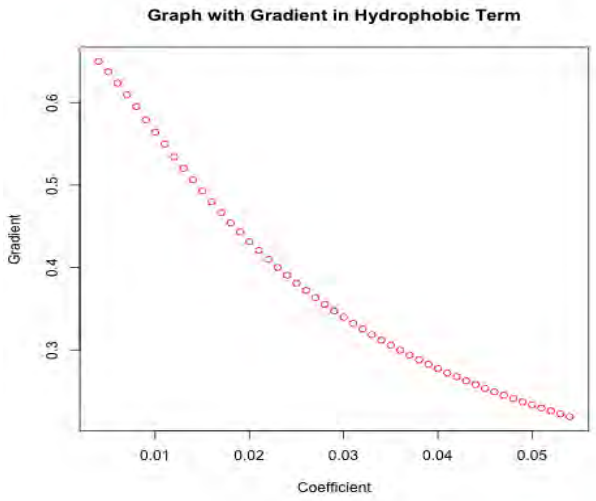

Figure 4.5: Plot of gradient vs coefficients for the hydrophobic terms

4.4 Re-parameterization of X-score

Improving a scoring function can be done either by adding a new term or just by re-parametrising the existing model to fit the users needs. As our primary protein family is the kinase protein family, our aim was to re-parameterised X-Score so that it can be used for this specific protein family.

As hydrogen bonding play a crucial role during the binding process, we split this term into two part : with one term for charged atoms and an another term for uncharged atoms.

Table 4.13 shows the charged amino acids that we have taken into account for the re-parametrisation process.

4.4.1 Ligands structures

Figures (4.7) ,(4.8) and (4.9) are ligand structures of adenosine triphosphate, adenosine diphosphate and imatinib bound to CDK2. The interactions shown in each figures are those mediated by hydrogen bonds and by hydrophobic contacts. Hydrogen bonds are indicated by dashed lines between the atoms involved while hydrophobic contact are represented by an arc with spaces radiating towards the ligands atoms they contact. The contacted atoms are shown with spokes radiating back [RCSB Protein Data

Table 4.13: Charged amino acids

Amino acid ID	Chemical properties	Physical propeties	Side chains	Name
Asp	Acidic	Polar (charged)	$OD1/OD2$	Aspartic Acid D
Glu	Acidic	Polar (charged)	$OE1/OE2$	Glutamic Acid E
Lys	Basic	Polar (positive charged)	NE	Lysine K
Arg	Basic	Polar (positive charged)	$NE/NH1/NH2$	Arginine R
His	Basic	Polar (positive charged)	NE	Histidine H

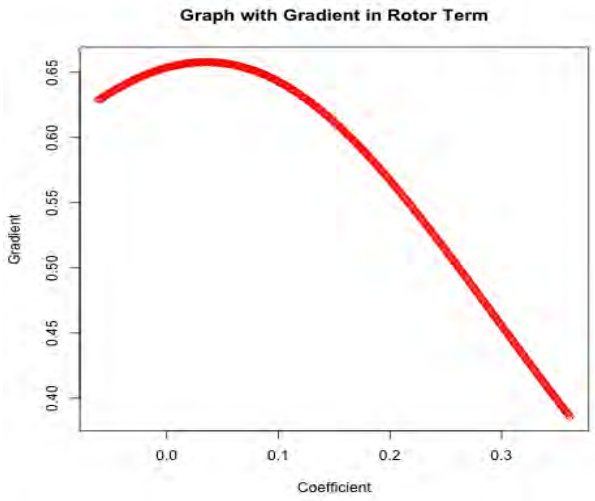

Figure 4.6: Plot of gradient vs coefficients for the rotor terms

Bank].

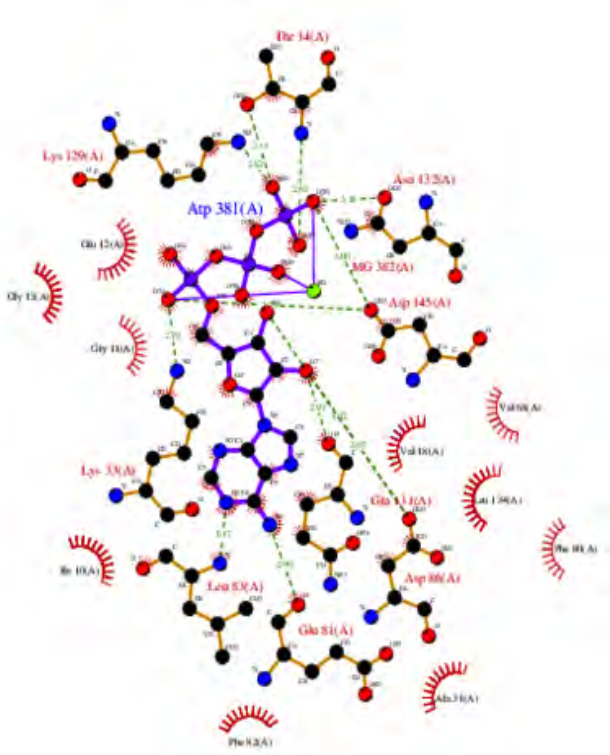

Figure 4.7: Ligand : structure of adenosine triphosphate ATP bound to CDK2

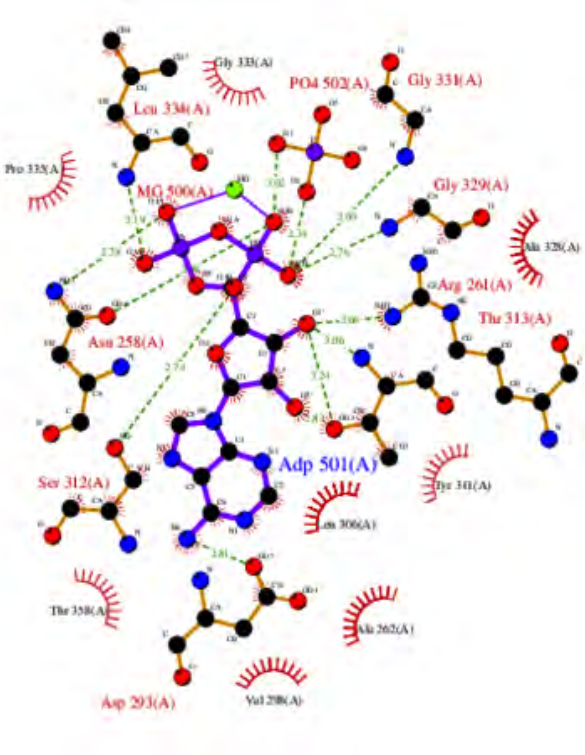

Figure 4.8: Ligand : structure of adenosine diphosphate (ADP) bound to CDK2

4.5 Re-pararameterisation model

We recall Equations (3.1),(3.2) and (3.3) each have four parameters namely the Van der Waal (VDW), hydrogen bonding (HB) and hydrophobic pair (HP)/ hydrophobic match (HM)/hydrophobic surface (HS). In our model we are going to break the hydrogen bonding into two terms. one term will be for charged amino acid side chains and the other term will be for the uncharged amino acid side chains.

If we denote for the hydrogen bonding for charged atoms by $H-Bond^c$ and for uncharged atoms $H-Bond^u$, the equations 3.1), (3.2) and (3.3) can now be written as

$$\begin{aligned} HPScore = C_{0,1} + C_{VDW,1} \times (VDW) + C_{HB,11} \times (H-Bond^c) \\ + C_{HB,12} \times (H-Bond^u) + C_{HP} \times (HP) + C_{R,1} \times (Rotor) \end{aligned} \tag{4.57}$$

$$\begin{aligned} HMScore = C_{0,2} + C_{VDW,2} \times (VDW) + C_{HB,21} \times (H-Bond^c) \\ + C_{HB,22} \times (H-Bond^u) + C_{HM} \times (HM) + C_{R,2} \times (Rotor) \end{aligned} \tag{4.58}$$

$$\begin{aligned} HSScore = C_{0,3} + C_{VDW,3} \times (VDW) + C_{HB,31} \times (H-Bond) \\ + C_{HB,32} \times (H-Bond^u) + C_{HS} \times (HS) + C_{R,3} \times (Rotor) \end{aligned} \tag{4.59}$$

Where coefficients $C_{0,1}$; $C_{VDW,1}$, $C_{HB,11}$, $C_{HB,12}$, C_{HP}, $C_{R,1}$ in HPScore, $C_{0,2}$, $C_{VDW,2}$, $C_{HB,21}$, $C_{HB,22}$, C_{HM}, $C_{R,2}$ for the HMScore, $C_{0,3}$, $C_{VDW,3}$, $C_{HB,31}$, $C_{HB,32}$, C_{HS}, $C_{R,3}$ for the HSScore can be found by multiple regression.

4.5.1 Least Square Regression

We made use of R software to compute all the coefficients above by multiple linear regression.

Hydrophobic pair based scoring function A summary of the results for the hydrophobic pair based-scoring function are shown in Table 4.14.

R gives more than one value for the coefficients of the uncharged terms but the coefficients for which the significance code was greater than 0.05 were discarded as they are not statistically significant. For those statistically significant we considered only the one that has the smallest $P-value$ i.e $C_{HB^u,14.0}$.

Hydrophobic match based scoring function A summary of the results for the hydrophobic match based-scoring function are shown in Table (4.15).

As for Hydrophobic pair the regression for Hydrophobic match yield multiple value for the uncharged term coefficient and the most significant one has a $P-value$ of 0.000524 i.e $C_{HB^u,24.0}$. This shows that the constant or intercept of our regression is not statistically significant as the $P-value$ is greater than 0.05.

Table 4.14: coefficients for Equation (4.57)

Coefficients	Estimates	Std Error	t value	P-value	Significance codes
$C_{0,1}$	0.0839528	0.0410622	2.045	0.041134	0.05
$C_{VDW,1}$	0.0045956	0.0001954	23.513	$<$ 2e-16	0.001
$C_{HB^c,11}$	0.0671145	0.0059993	11.187	$<$ 2e-16	0.001
$C_{HB^u,128}$	0.1604596	0.0577827	2.777	0.005578	0.01
$C_{HB^u,12.0}$	0.0345801	0.0408915	0.846	0.397925	1
$C_{HB^u,12.1}$	0.0291940	0.0441360	0.661	0.508455	1
$C_{HB^u,12.2}$	0.0721876	0.0577837	1.249	0.211825	1
$C_{HB^u,12.3}$	0.0454549	0.0456834	0.995	0.319951	1
$C_{HB^u,12.4}$	0.0490832	0.0500597	0.980	0.327053	1
$C_{HB^u,12.5}$	0.0715533	0.0500415	1.430	0.153029	1
$C_{HB^u,12.6}$	0.1433458	0.0577939	2.480	0.013273	0.05
$C_{HB^u,12.7}$	0.0794044	0.0447587	1.774	0.076324	0.1
$C_{HB^u,12.8}$	0.0979413	0.0577903	1.695	0.090395	0.1
$C_{HB^u,12.9}$	0.0919027	0.0500444	1.836	0.066559	0.1
$C_{HB^u,13.0}$	0.1086949	0.0418279	2.599	0.009482	0.01
$C_{HB^u,13.7}$	0.1610293	0.0577846	2.787	0.005414	0.01
$C_{HB^u,13.8}$	0.1658454	0.0578028	2.869	0.004193	0.01
$C_{HB^u,14.0}$	0.1742004	0.0500434	3.481	0.000519	0.001
C_{HP}	0.0095084	0.0002867	33.170	$<$ 2e-16	0.001
$C_{R,1}$	-0,0839061	0.0052379	-16,019	$<$ 2e-16	0.001

Table 4.15: coefficients for Equation (4.58)

Coefficients	Estimates	Std Error	t value	P-value	Significance codes
$C_{0,2}$	0.070672	0.040734	1.735	0.083024	0.1
$C_{VDW,2}$	0.005231	0.000190	27.530	$<$ 2e-16	0.001
$C_{HB^c,21}$	0.066979	0.005952	11.254	$<$ 2e-16	0.001
$C_{HB^u,228}$	0.160523	0.057333	2.800	0.005200	0.01
$C_{HB^u,22.0}$	0.036499	0.040572	0.900	0.368512	1
$C_{HB^u,22.1}$	0.027396	0.043792	0.626	0.531711	1
$C_{HB^u,22.2}$	0.071107	0.057333	1.240	0.215145	1
$C_{HB^u,22.3}$	0.046901	0.045327	1.035	0.301028	1
$C_{HB^u,22.4}$	0.048494	0.049670	0.976	0.329110	1
$C_{HB^u,22.5}$	0.071077	0.049652	1.432	0.152562	1
$C_{HB^u,22.6}$	0.139660	0.057343	2.436	0.015025	0.05
$C_{HB^u,22.7}$	0.078769	0.044410	1.774	0.076387	0.1
$C_{HB^u,22.8}$	0.094891	0.057340	1.655	0.098225	0.1
$C_{HB^u,22.9}$	0.090092	0.049654	1.814	0.069886	0.1
$C_{HB^u,23.0}$	0.106329	0.041502	2.562	0.010535	0.05
$C_{HB^u,23.7}$	0.162554	0.057334	2.835	0.004662	0.01
$C_{HB^u,23.8}$	0.170802	0.057352	2.978	0.002962	0.01
$C_{HB^u,24.0}$	0.172707	0.049653	3.478	0.000524	0.001
C_{HM}	0.267259	0.007932	33.696	$<$ 2e-16	0.001
$C_{R,2}$	-0.085362	0.005189	-16.451	$<$ 2e-16	0.001

Table 4.16: coefficients for Equation (4.59)

Coefficients	Estimates	Std Error	t value	P-value	Significance codes
$C_{0,3}$	0.0694501	0.0414868	1.674	0.094403	0.1
$C_{VDW,3}$	0.0052895	0.0001933	27.359	$<$ 2e-16	0.001
$C_{HB,31}$	0.0627287	0.0060501	10.368	$<$ 2e-16	0.001
$C_{HB^u,328}$	0.1605290	0.0583923	2.749	0.006071	0.01
$C_{HB^u,32.0}$	0.0415294	0.0413190	1.005	0.315069	1
$C_{HB^u,32.1}$	0.0273010	0.0446015	0.612	0.540588	1
$C_{HB^u,32.2}$	0.0710079	0.0583932	1.216	0.224228	1
$C_{HB^u,32.3}$	0.0470336	0.0461652	1.019	0.308513	1
$C_{HB^u,32.4}$	0.0491235	0.0505878	0.971	0.331729	1
$C_{HB^u,32.5}$	0.0710329	0.0505694	1.405	0.160398	1
$C_{HB^u,32.6}$	0.1393211	0.0584030	2.386	0.017220	0.05
$C_{HB^u,32.7}$	0.0787105	0.0452309	1.740	0.082098	0.1
$C_{HB^u,32.8}$	0.0946105	0.0583997	1.620	0.105501	1
$C_{HB^u,32.9}$	0.0899250	0.0505722	1.778	0.075649	0.1
$C_{HB^u,33.0}$	0.1063132	0.0422690	2.515	0.012036	0.05
$C_{HB^u,33.7}$	0.1626947	0.0583941	2.786	0.005423	0.01
$C_{HB^u,33.8}$	0.1712579	0.0584118	2.932	0.003437	0.01
$C_{HB^u,34.0}$	0.1725697	0.0505712	3.412	0.000667	0.001
C_{HS}	0.0031169	0.0000960	32.466	$<$ 2e-16	0.001
$C_{R,3}$	-0.0882309	0.0052737	-16.731	$<$ 2e-16	0.001

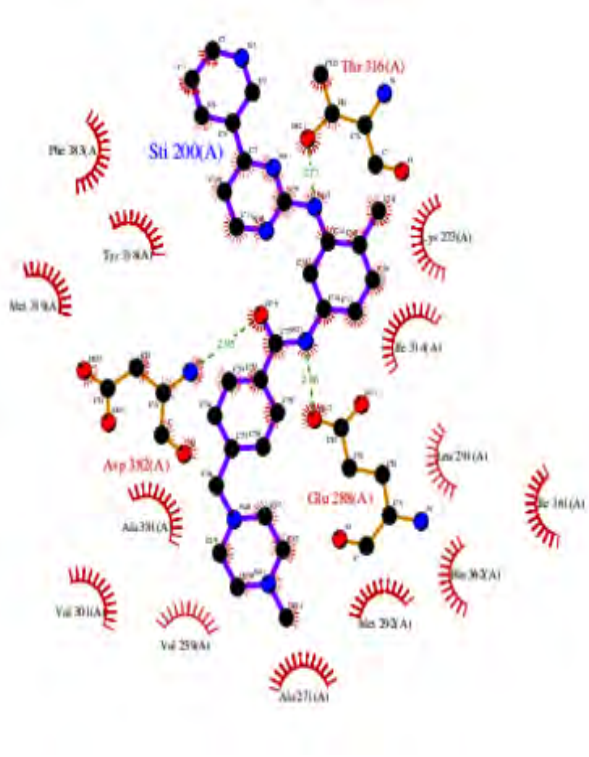

Figure 4.9: Ligand : structure of Imatinib or STI bound to CDK2

Hydrophobic surface based scoring function The values for the hydrophobic surface term coefficients for equation (4.59) are shown in Table 4.16.

4.5.2 Statistical analysis

A statistical analysis of the three scoring functions is shown in Table 4.17

Table 4.17 shows that both R and adjusted R-squared are greater than 0.70 meaning that more than 70% of our data is explained by the multiple regression model. This is a much higher proportion of the data accounted for by than the original X-score scoring functions. F-statistic help us to validate the

Table 4.17: Statistical summary

Statistics	Hydrophobic Pair	Hydrophobic Match	Hydrophobic Surface
Multiple R-squared	0.7234	0.7277	0.7175
Adjusted R-squared	0.7187	0.7231	0.7128
F-statistic	155.1	158.5	150.7
Residual standard error	0.04086	0.04054	0.04129
P-value	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
degrees of freedom (DF)	1127	1127	1127

Table 4.18: Test data

Complexes	**Description**	**Experimental** pK_d	**X-Score** pK_d
2J6M/AEE	EGFR/AEE788	7.56	6.95
2JIU/AEE	EGFR/AEE788	7.56	7.18
2ITT/AEE	EGFR(L858R mutation)/AEE788	7.73	7.16
2ITP/AEE	EGFR(G719S mutation)/AEE788	7.95	7.17
1M17/AQ4	EGFR/Erlotinib	8.80	5.75
2ITY/IRE	EGFR/Iressa	6.91	5.91
2ITO/IRE	EGFR/Iressa	6.91	6.02
2ITZ/IRE	EGFR/Iressa	7.96	6.06
2ITU/STU	EGFR(L858R)/Staurosporine	7.15	6.34
2ITW/STU	EGFR/Staurosporine	7.15	6.32
2ITQ/STU	EGFR(G719S)/Staurosporine	7.15	6.42
1XKK/FMM	EGFR/Quinazoline inhibitor	8.07	8.07

model as the calculated value show here is far greater than the critical value of 19 and 1127 degrees of freedom.

4.6 Prediction

In this section we test our re-parameterised X-Score model against 12 Epidermal growth factor receptor (EGFR) to see how well it performs compared to the original X-Score. The Table 4.18 shows all the data used for calculation. The Figure 4.10 shows the scatter plot and correlation between the experimental and calculated pK_d values of 12 EGFR complexes in the test data set.

The computation of the regression line from the data in Table 4.18 yield the equation

$$Y = 5.4683 + 0.1603X \tag{4.60}$$

Where Y denote the X-Score predicted pK_d while represents X the observed or experimental pK_d and the coefficient of determination $R^2 = 0.01653$ showing a poor goodness of fit (i.e only 1.65 of our data is explained by the regression line) the p-value here is 0.6505. Based on this high p-value at the level of 95% confidence interval, points to the fact that the correlation between the experimental pK_d and the X-Score pK_d for the test data occurred by chance. This is understandable as we emphasize in Chapters 2 and 3 that X-Score is an empirical scoring function as such it is size dependent. The test data used here consist merely of 12 sets of data. An other reason for this poor performance of X-Score on epidermal growth factor receptor is that X-Score failed to capture the energetic penalty due to the effect of mutation (e.g on the enzyme). We will suggest some path of future investigation on this matter in the next section.

The Table 4.19 shows the experimental pK_d, the X-Score pK_d and our model pK_d, i.e the weighted Scoring function pK_d. To compute the pK_d in our model we make use of the equations (4.57), (4.58) and (4.59) where in the Hydrogen bonding we distinguish between the charged and uncharged terms.

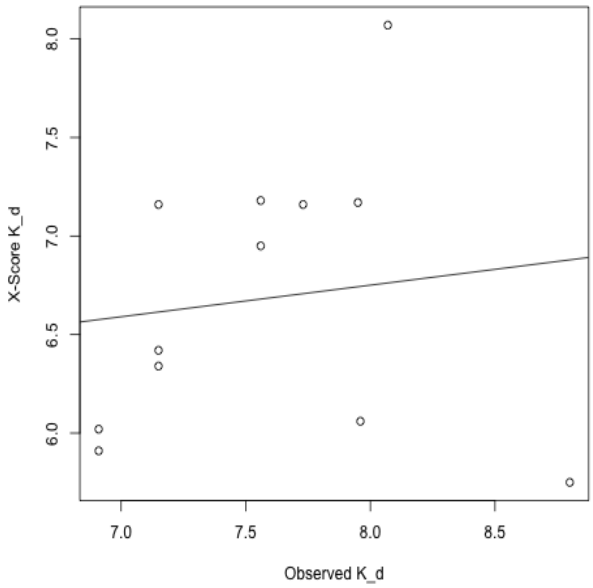

Figure 4.10: Correlation between the experimental and X-Score pK_d values of 12 EGFR complexes in the test set.

Table 4.19: pK_d for Weighted Scoring function vs Experimental pK_d and X-Score pK_d

PDB/Ligand ID	**Experimental** pK_d	**X-Score** pK_d	**Re-parametrised X-Score** pK_d
2J6M/AEE	7.56	6.95	7.32
2JIU/AEE	7.56	7.18	7.66
2ITT/AEE	7.73	7.16	7.58
2ITP/AEE	7.95	7.17	7.62
2ITY/IRE	6.91	5.91	6.36
2ITO/IRE	6.91	6.02	6.52
2ITZ/IRE	7,96	6.06	6.53
2ITU/STU	7.15	6.34	6.69
2ITW/STU	7.15	6.42	6.74
2ITQ/STU	7.15	7.16	7.77
1M17/AQ4	8.80	5.75	7.27
1XKK/FMM	8.07	8.07	7.79

The data in Table 4.19 yield the equation

$$Y = 2.8834 + 0.5748X \tag{4.61}$$

as the regression fitting line between the experimental pK_d and the computed pK_d from the weighted scoring function; the coefficient of determination $R^2 = 0.2102696$, suggesting that the goodness of fit is now 21% compared to 1.65% obtained by X-Score. In our model the p-value is 0.1338 even though this value indicates that the result is not statistically significant at the level of 95% confidence interval but the trend has improved. Figure 4.11 shows the correlation between the experimental and weighted pK_d for 12 EGFR complexes in the test data set.

Since all the above tests resulted in non statistically significant correlation, we went further to test whether the difference in the correlation is statistically significant, and we found that the probability that there is no difference between the two correlations of determination is 0.00003 meaning that it is unlikely that this difference is due to chance, even though the re-parametrised coefficients were obtained for analysis of ligand bound structures of different kinase, CDK2. This appears to bode well for the future applicability of our heuristic approach.

4.7 Future work

4.7.1 Effect of mutation on protein structure

When calculating pK_d values, X-Score only consider the ligand-bound complex and then assumes a fixed average relationship between number of atomic interaction and free energy of binding. Thus, if a mutation affects the stability of the free and ligand-bound protein equally, X-Score should compute a pK_d value equally well for the mutant protein, but will not do so if the stability effect is differential on free and ligand-bound protein. However, it is possible that such differential effects may be calculable, in which case they could be added as additional terms in X-Score.

For example suppose both the wild type (wt) and mutant protein have the same number of pairwise interaction with ligand but the mutation causes subtle difference in bond length as well as destabilising the protein structure. This situation can be written by the equation

$$E_u + L \rightleftharpoons E_f + L \rightleftharpoons E_f.L$$

Where E_u and E_f is protein or enzyme in unfolded and folded state respectively and L is the ligand.

In Figure 4.12, X-Score can predict the quantity in A which is $\triangle G$ (formation of wt-ligand complex) and we can also find the quantity A from experimental data. Protein stability prediction algorithm such as SDM can predict the quantity in B which is $\triangle G$ (wt-mutant), i.e the free energy penalty of mutation on protein stability.

If we could calculate the quantity in C in the Figure 4.12 by similar means (where C equal the free energy penalty of mutation on the stability of the protein-ligand complex), then we could compute the quantity in D (where D is the free energy charge on binding of a ligand to the mutant protein).

We now have the following,

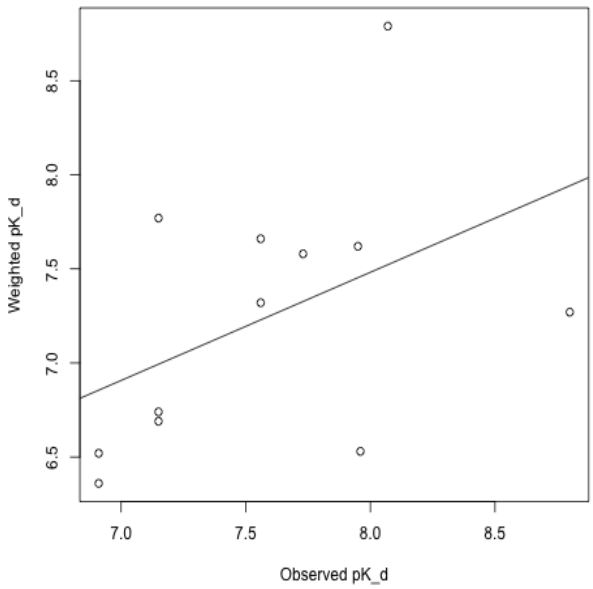

Figure 4.11: Correlation between the experimental and weighted pK_d values of 12 EGFR complexes in the test set.

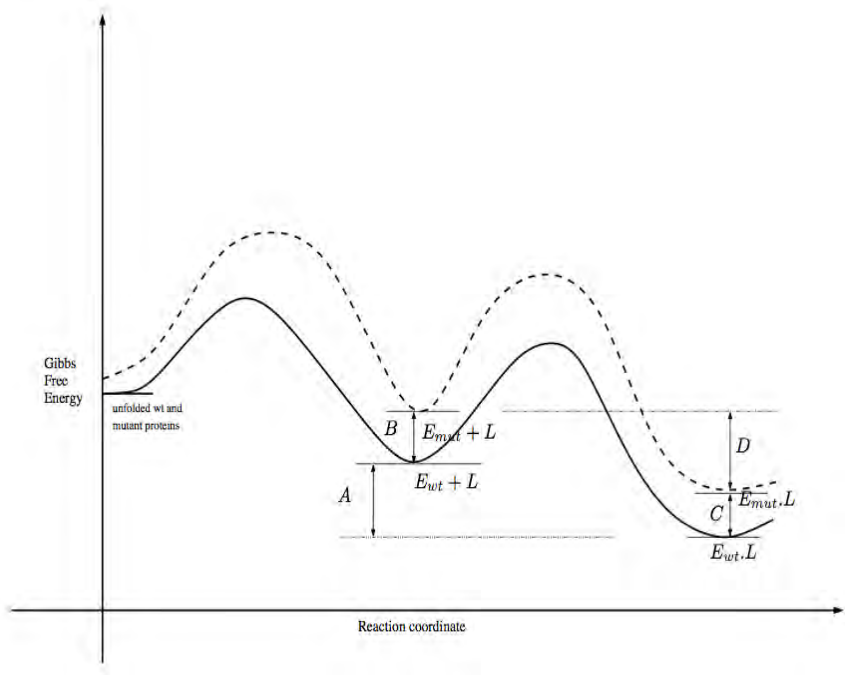

Figure 4.12: Effect of mutation on protein stability

$A = \triangle G_{(wt,true)}$ = calculated from observed K_d,

$D = \triangle G_{(mut,true)}$ = A + B - C

We can estimate the quantity B from a software like SDM ($= G_{wt} - G_{mut}$ $(= \triangle G_{SDM})$. Then let define $C = \triangle G_{prot} + \triangle G_{prot.ligand}$ where $\triangle G_{prot}$ equal the energy penalty of mutation on protein stability in the ligand-bound state and $\triangle G_{prot.ligand}$ equal the energy penalty of mutation on protein-ligand interaction.

If we assume that

$$\triangle G_{prot-Ligand} = \triangle G_{(wt\,,X-Score)} - \triangle G_{(mut\,,X-Score)}$$

here we use X-Score to tell us about the altered ligand-bound environment in the mutant.

Therefore,

$$\triangle G_{(mut\,,true)} = \triangle G_{(wt\,,true)} + \triangle G_{SDM} - \triangle G_{prot} - (\triangle G_{(wt\,,X-Score)} - \triangle G_{(mut\,,X-Score)})$$

or simply as

$$\triangle G_{(mut\,,true)} = \left[\triangle G_{(wt\,,true)} - \triangle G_{(wt\,,X-Score)}\right] + \left[\triangle G_{SDM} - \triangle G_{prot}\right] + \triangle G_{(mut\,,X-Score)}$$

where the quantity $\left[\triangle G_{(wt\,,true)} - \triangle G_{(wt\,,X-Score)}\right]$ is a general correction on X-Score and $\left[\triangle G_{SDM} - \triangle G_{prot}\right]$ define the specific correction for differential effect of mutation on the protein stability in the free and ligand-bound forms.

Therefore in principle we just need to calculate the energy penalty of mutation in wild type (wt) ligand-free protein (i.e $\triangle G_{SDM}$) and a wt ligand-bound protein (i.e $\triangle G_{prot}$) and we can then accurately compute the free energy of binding of the ligand to mutant.

One of the question one can ask is should we really consider all ligand binding free energies relative to the unfolded protein state ?

In the other hand we know from

$$\triangle G = -RT\,Log\,K_a \tag{4.62}$$

If $\triangle G$ is negative then $K_a > 1$ and interaction is favourable where K_a is defined as

$$K_a = \frac{[E_f.L]}{[E_f] \times [L]}$$

and

$$K_d = \frac{[E_f] \times [L]}{[E_f.L]}$$

X-score computes the binding free energy $\triangle G$ by subtracting the energy in the unbound state from the bound state meaning when the binding is favourable i.e $\triangle G$ is negative. This free energy is expressed by the equation

$$\triangle G = [G_{E.L} - G_{E+L}] \tag{4.63}$$

And the energetic penalty of mutation on the stability of the bound complex is given by

$$\triangle\triangle G = \triangle G_{wt} - \triangle G_{Mut} \tag{4.64}$$

This equation gives the energetic effect of mutation on the protein and this effect may be neutral or destabilising if $\triangle\triangle G$ is negative.

Likewise to calculate the effect of a mutation on a ligand bound complex X-Score would need to be modified by the calculated factor above.

We would need to investigate these 2 hypothesis:

1. Hypothesis 1 : Should we really be considering all ligand binding free energy relatives to the unfolded protein state ?
2. Hypothesis 2 : is there some approximate proportional relationships between $\triangle G_{SDM}$ and $\triangle\triangle G$?

where $\triangle G_{SDM}$ represents the energetic penalty between the free wild type and free mutant in the folded state.

4.7.2 User interface for scoring function for protein family

Another path of investigation will be to develop software tools that allow researchers to select and modulate the calculated contribution of sets of amino acid residues when applying generalist scoring functions to drug target from specific protein families. Once this user interface has been developed the next task will be to test the utility of these tools on the X-score scoring function and EGFR kinase - a drug target in Non Small Cell Lung Cancer (NSCLC). Any given scoring function would require its own interface, but once developed, such an interface would be usable for any protein family.

To achieve this a variety of strategies may be applied to select these sets of residues. In the case of the EGFR kinase protein family a promising initial strategy in the definition of sets of residues to target for modulation will be structure alignments of variants of the protein, with initial weightings based thereafter on hydrophobic/hydrophilic character considerations, together with information on conformational changes that occur on ligand binding. This is a general strategy that may be applied to any protein family where structures of at least two variants are available.

5. Conclusion

In this thesis we presented a comprehensive review of Scoring functions as a method to compute the free binding energy between the protein and the drug-like molecules called ligand. We choose X-Score scoring function as our case study in our re-visitation and re-computation of scoring function. We addressed the question of sensitivity of X-Score by asking how a coefficient (or weight) of a term in X-Score vary when the original coefficient value undergoes a small perturbation. We tested the sensitivity of X-Score on the target training data (data in Table (4.7) the X-Score shows some robustness when the coefficients of hydrogen bonding term undergoes small perturbation the strength of goodness of fit is inversely proportional to the change coefficients of the terms in X-Score. A notable sensitivity was observed in the weighting of the hydrophobic terms perturbation. The highest strength of correlation were obtained when the coefficient of determination was 30.58%. And when the gradient reaches its optimal value, the coefficient of determination is 24.75% meaning that as the gradient decreases from 0.650051 to 0.305819, the proportion of data explained by the fitting line increased by approximately 5%.

There are many strategies to adopt in order to improve a scoring function and this can be done either by adding new terms or by re-paramatizing the existing weighted. In our study we adopted the latter by splitting the hydrogen bonding term into two different terms, we have a term for charged and uncharged amino acid in the ligand. By doing so we obtained through regressional analysis all the coefficients of all terms with a multiple R-squared equal to 0.7234 (or 0.7187 for adjusted R-square) for which we deduce that 71% to 72% our data is explained by the model with a p-value almost zero we can say that this performance is not due to chance.

The test of the model shows a net improvement compared to original X-Score the result obtained was statistically meaningful due at level of 95% confidence interval. We constructed our model on an empirical scoring function as such it remains dependent to the size of the data. So we need more data set to test and and find out whether the improvement we found from the model is not due to chance.

Finally, we show two paths to further the investigation, in order to improve our heuristic method of weighted scoring function, one path is to see how the mutation affect the stability of the protein when computing the free binding energy and the second path is to design a user interface for scoring function family that will allow researchers to select and modulate the calculated contribution of sets of amino acid residues when applying commonly used scoring function.

Acknowledgements

The realization of this thesis would not have been possible without the guidance of my supervisor, help from friends, and support from my family and wife.

I would like to express my deepest gratitude to my my advisor, professor Jonathan Blackburn, for his outstanding guidance, financial support, patience, and providing me with an excellent atmosphere for doing research. I would also like to thank Dr. Gaston Mazandu for excellent assistance in programming. Special thanks goes to miss Lauren Coulson for proofreading the text and providing help with the best suggestions. I would like to thank the generous financial support from various sources such as The African Institute for Mathematical Sciences (AIMS), University of Cape Town for the award of International scholarships for postgraduate student, Faculty of Health Sciences's departmental grant.

Finally, I would like to thank my wife, Josephine Mangaza Kisua for always being their for me and stood by me through good and bad times. For that I would like to dedicate this thesis to my son Mambo Hilaire, Jr.

References

[Wan, n.d.]

[Ajay & Murcko, 1995] Ajay, & Murcko, MA. 1995. Computational methods to predict bindig free energy in ligand-receptor complexes. *J. Med. Chem.*, 1684 – 1657.

[Andrews *et al.*, 1984] Andrews, PR., Craik, DJ., & Martin, JL. 1984. Functional-group contributions to drug receptor interactions. *J. Med. Chem.*, 1648 – 1657.

[Atkins, 1998] Atkins, PW. 1998. *Physical chemistry.6th ed.* Oxford University Press.

[Bohm, 1994] Bohm, HJ. 1994. The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known 3-dimensional structure. *J. Comput. Aided Mol. Des.*, 243 – 256.

[Bohm & Klebe, 1996] Bohm, HJ., & Klebe, G. 1996. What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs? *Angew. Chem. Ed. Engl.*, 2614–2614.

[Bostrom *et al.*, 1998] Bostrom, J., Norrby, PO., & Liljefors, T. 1998. Conformational energy penalties of protein-bound ligands. *J. Comput. Aided Mol. Des.*, 383 – 396.

[Brooks *et al.*, 1983] Brooks, BR., Bruccoleri, RE., Olafson, B.D, States, DJ., Swaminathan, S., & Karplus, M. 1983. a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 187 – 217.

[Charifson *et al.*, 1999] Charifson, PS., Corkery, JJ., Murcko, MA., & Walters, WP. 1999. Consensus scoring:A method for obtaining improved hit rates from docking databases of three dimensional structures into proteins. *J. Med. Chem.*, 5100 – 5109.

[Coetzer, 2011] Coetzer, N. (ed). 2011. *Statistics, Lecture Notes*. SANBI.

[Cornell *et al.*, 1995] Cornell, WD., Cieplak, P., Bayly, CI., Gould, IR., Merz, KM., Ferguson, DM., Spellmeyer, WDC., Fox, T., Caldwell, JW., & Kollman, PA. 1995. A 2nd generation force-field for the simulation of proteins, nucleic-acids, ands organic-molecules. *J. Am. Chem. Soc.*, 337 – 344.

[DeWitte & Shakhnovich, 1996] DeWitte, RS., & Shakhnovich, EI. 1996. SMoG: De Novo design method based on simple, fast, and accurate free energy estimates.1. Methodology and supporting evidence. *J. Am. Chem. Soc.*, 11733 – 11744.

[DeWitte *et al.*, 1997] DeWitte, RS., Ishchenko, AV., & Shakhnovich, EI. 1997. SMoG: De novo design method based on simple, fast, and accurate free energy estimates.2. Case studies in molecular design. *J. Am. Chem. Soc.*, 4608 – 4617.

[Draper & Smith, 1998] Draper, NR., & Smith, H. 1998. *Applied Regression Analysis, 3rd ed.* John Willey & Sons, Inc.

[Eldridge *et al.*, 1997] Eldridge, MD., Murray, CW., Auton, TR., Paolini, GV., & Mee1, RP. 1997. Empirical scoring functions.1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol.*, 425 – 445.

[Evans & McLeod, 2003] Evans, WE., & McLeod, ML. 2003. Pharmacogenomics - Drug disposition, drug targets, and side effects. *New Engl. J. Med*, 538–549.

[Fenu *et al.*, 2007] Fenu, A. Luca, Lewis, A. Richard, Good, C. Andrew, Bodkin, M., & Essex, JW. 2007. Structure-based Drug Discovery : From Free-energies of Binding to Enrichment in Virtual screening. *Springer*, 223 – 245.

[Ferara *et al.*, 2004] Ferara, P., Gohlke, H., Price, D., Klebe, G., & Brooks, CL. 2004. *J. Med. Chem.*, 3032 – 3047.

[Friesner *et al.*, 2004] Friesner, RA., Banks, JL., Murphy, RB., Halgren, TA., Klicic, JJ., Mainz, DT., Repasky, MP., Knoll, EH., Shelley, M., & Perry., JK. 2004. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, 1739 – 1749.

[Gehlhaar *et al.*, 1995] Gehlhaar, DK., Verkhivkerl, GM., Rejtol, PA., & Freer, ST. 1995. Molecular recognition of the inhibitor Ag-1343 by Hiv-1 Protease - conformationally flexible docking by evolutionary programming. *Chem. Biol.*, 317 – 324.

[Giordanetto *et al.*, 2004] Giordanetto, F., Cotesta, S., Catana, C., Trosset, JY., Vulpetti, A., Stouten, PFW., & Kroemer, RT. 2004. Novel scoring functions comprising QXP, SASA, and protein side-chain entropy terms. *J. Chem. Inf. Comput. Sci.*, 882 – 893.

[Gohlke *et al.*, 2000] Gohlke, H., Hendlich, M., & Klebe, G. 2000. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 337–356.

[Guo *et al.*, 2004] Guo, JX., Hurley, MM., Wright, JB., & Lushington, GH. 2004. A docking scoring function for estimating ligand-protein interactions: Application to acetylcholinesterase inhibition. *J. Med. Chem.*, 5492 – 5500.

[Halgren, 1996] Halgren, TA. 1996. Merck Molecular Force Field. *Journal of Computational Chemistry*, **17**, 490 – 519.

[Halgren *et al.*, 2004] Halgren, TA., Murphy, RB., Friesner, RA., Beard, HS., LL. Frye, WT. Pollard, & Banks, JL. 2004. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.*, 1750 – 1759.

[Halperin *et al.*, 2002] Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins: Structure, Function, and Genetics. *Wisley-Lyss inc.*, 409 – 443.

[Haupt & Haupt, 2004] Haupt, RL., & Haupt, SE. 2004. *Pratical Genetic Algorithms, Second Edition*. John Willey & Sons, Inc.

[Hu *et al.*, 2005] Hu, L., Benson, ML., Smith, RD., Lerner, MG., & Carlson, HA. 2005. Binding MOAD (Mother of All Databases). *Proteins*, 333 – 40.

[Ishchenko & Shakhnovich, 2001] Ishchenko, AV., & Shakhnovich, EI. 2001. SMall molecule growth (SMoG2001): An improved knowledge-based scoring function for protein-ligand interactions. *J. Med.Chem.*, 2770 – 2780.

[Jacobsson *et al.*, 2003] Jacobsson, M., Lidé, P., Stjernschantz, E., Bonström, H., & Norinder, UJ. 2003. Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.*, 5781 – 5789.

[Jhoti & Leach, 2007] Jhoti, H., & Leach, AR. (eds). 2007. *Structure - Based Drug Discovery*. Springer.

[Jones *et al.*, 1997] Jones, G., Willett, P., Glen, RC., Leach, AR., & Taylor, R. 1997. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 727 – 748.

[Kellenberge *et al.*, 2004] Kellenberge, G., Rodrigo, J., Muller, P., & Rognan, D. 2004. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Willey-Lyss inc.*, 225 – 242.

[Klon *et al.*, 2004] Klon, AE., Glick, M., Thoma, M., Acklin, P., & Davies, JW. 2004. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J. Med. Chem.*, 2743 – 2749.

[Kollman, 1993] Kollman, P. 1993. Free-energy calculations - applications to chemical and biochemical phenomena. *Chem. Rev*, 2395 – 2417.

[Kontoyianni *et al.*, 2004] Kontoyianni, M., McClellan, LM., & Sokol, GS. 2004. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.*, 558 – 565.

[Krovat & Langer, 2004] Krovat, EM., & Langer, T. 2004. Impact of scoring function on enrichment in docking-based virtual screening: An application study on rein inhibitors. *J. Chem. Inf. Comput. Sci.*, 1123 – 1129.

[Lipinski *et al.*, 2001] Lipinski, CA., Lombardo, F., Dominy, BW., & Feeney, PJ. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development setting. *Adv. Drug. Deliv. Rev.*, 3 – 26.

[Liu *et al.*, n.d.] Liu, T., Lin, Y., Wen, X., Jorrisen, RN., & Gilson, MK. *BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities Nucleic Acids Research*. http://www.bindingdb.org.

[Mancera *et al.*, 2004] Mancera, RL., Kallblad, P., & Todorov, NP. 2004. Ligand-protein docking using a quantum stochastic tunneling optimization method. *J. Comput. Chem.*, 858 – 864..

[Mark & Gusteren, 1994] Mark, AE., & Gusteren, WF. 1994. Decomposition of free-energy of a system in terms of specific interactions - implications for theoretical and experimental studies. *J. Mol. Biol.*, 167–176.

[Mitchell *et al.*, 1999] Mitchell, JBO., Laskowski, RA., Alex, A., & Thornton, JM. 1999. BLEEP - Potential of mean force describing protein-ligand interactions: II. Calculation of binding energies and comparison with experimental data. *J. Comput. Chem.*, 1177 – 1185.

[Mizutani *et al.*, 1994] Mizutani, MY., Tomioka, N., & Itai, A. 1994. Rational automatic search method for stable docking models of protein and ligand. *J. Mol. Biol.*, 310 – 326.

[Muegge & Martin, 1999] Muegge, IYC., & Martin. 1999. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.*, 791 – 804.

[Paez, 2004] Paez, JG. 2004. EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. *Science*, 304.

[Pan *et al.*, 2003] Pan, YP., Huang, N., Cho, S., MacKerell, AD., & Jr. 2003. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.*, 267 – 272.

[Pearlman, 1999] Pearlman, DA. 1999. Free energy grids: a practical qualitative application of free energy perturbation to ligand design using OWFEG method. *J. Med. Chem.*, 4313 – 4324.

[Pearlman & Charifson, 2001] Pearlman, DA., & Charifson, PS. 2001. Improved scoring of ligand-protein interactions using OWFEG Free energy grids. *J. Med. Chem.*, 502 – 511.

[Perola *et al.*, 2004] Perola, E., Walters, WP., & Charifson, PS. 2004. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins: Structure, Function, and Bioinformatics. *Wisley-Lyss, inc.*, 235 – 249.

[R Development Core Team, 2011] R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Raha, 2005] Raha, K. 2005. Pairwise decomposition of residue interaction energies using semi-empirical quantum mechanical methods in studies of protein-ligand interaction. *J. Am. Chem. Soc.*, 6583 – 6594.

[Raha & Merz, 2004a] Raha, K., & Merz, KM. 2004a. A quantum mechanics-based scoring function: study of zinc ion-mediated ligand binding. *J. Am. Chem. Soc.*, 235–249.

[Raha & Merz, 2004b] Raha, K., & Merz, KM. 2004b. Zinc mediated ligand binding: a quantum mechanics based approach. *J. Am. Chem. Soc.*, 1016–1017.

[Raha & Merz, 2005] Raha, K., & Merz, KM. 2005. Large -scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding binding mode of a diverse set of protein-ligand complexes. *J. Med. Chem.*, 4558 – 4575.

[Roche *et al.*, 2001] Roche, O., Kiyama, R., & Brooks, CL. 2001. Ligand - protein database: Linking protein-ligand complex structures to binding data. *J. Med. Chem.*, 392 – 3598.

[Shoichet *et al.*, 1999] Shoichet, BK., Leach, AR., & Kuntz, ID. 1999. Ligand solvation in molecular docking. Proteins: Structure, Function, and Genetics. 4 – 16.

[Smith *et al.*, 2003] Smith, R., Hubbard, RE., Gschwend, DA., Leach, AR., & Good, AC. 2003. Analysis and optimization of structure-based virtual screening protocols (3). New methods and old problems in scoring function design. *J. Mol. Graph. Model.*, 41 – 53.

[Society, 2005] Society, Royal. 2005. Personalised medicines: hopes and realities. *Royal Socity Publishing*.

[Stahl & Rarey, 2001] Stahl, M., & Rarey, M. 2001. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.*, 1035 – 1042.

[Todorov *et al.*, 2002] Todorov, NP., Mancera, RL., Kallblad, P., & Monthoux, P. 2002. EasyDock: A new docking program for high-throughput screening and binding-mode search. *J. Am. Chem. Soc.*, 224.

[Verdonk *et al.*, 2003] Verdonk, ML., Cole, JC., Hartshorn, M., Murray, C., & Taylor, RD. 2003. Improved protein-ligand docking using Gold. Proteins: Structure, Function, and Genetics. *J. Med. Chem.*, 609 – 623.

[Verkhivker *et al.*, 1995] Verkhivker, G., Appelt, K., Freer, ST., & Villafranca, JE. 1995. Empirical free-energy calculations of ligand-protein crystallographic complexes.1. Knowledge-based ligand-protein interaction potentials applied to the prediction of human-immuno deficiency virus-1 protease binding-affinity. *Protein Eng.*, 677 – 691.

[Vieth *et al.*, 1998] Vieth, M., Hirst, JD., Kolinski, A., & Brooks, CL. 1998. Assessing energy functions for flexible docking. *J. Comput. Chem.*, 80 – 89.

[Vigers & Rizzi, 2004] Vigers, GPA., & Rizzi, JP. 2004. Multiple active site corrections for docking and virtual screening. *J. Med. Chem.*, 80 – 89.

[Wang *et al.*, 1998] Wang, R., Liu, L., Lai, L., & Tang, Y. 1998. SCORE : A New Empirical Method for Estimating the Binding Affinity of a protein-Ligand Complex. *J. Mol. Model*, **4**, 379 – 394.

[Wang & Wang, 2001.] Wang, RX., & Wang, SM. 2001.. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.*, 1422 – 1426.

[Wang *et al.*, 2004] Wang, RX., Fang, X., Lu, Y., & Wang, S. 2004. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, 2977–2980.

[Waters & MacLeod, 2003] Waters, JV., & MacLeod, HL. 2003. Cancer pharmacogenomics: current and future applications. *Biochim. Biophys.*, 99–111.

[Yang & Chen, 2001] Yang, JM., & Chen, CC. 2001. GEMDOCK: A generic evolutionary method for molecular docking. Proteins: Structure, Function, and Bioinformatics. *Wisley-Lyss inc.*, 288 – 304.

[Yun *et al.*, 2007] Yun, CH., Boggon, TJ., Woo, MS., Meyerson, M., & Eck, MJ. 2007. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: Mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell*, 217–227.

[Zou *et al.*, 1999] Zou, XQ., Sun, YX., & Kuntz, ID. 1999. Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. *J. Am. Chem. Soc.*, 8033 – 8043.